통계자료분석실습

UN DATA

2조 박상희 | 이인풍 | 김문경 | 정중한

2018. 11. 08

목요일 13:00

00 분석에 앞서..

UN 데이터란?

2009년 ~ 2011년 사이의 UN에 가입되어있는 국가들의 통계자료를 기록한 데이터이다.

• 어떤 국가들이 포함되어 있는가?

- 213개 국가의 국민 건강, 복지 및 교육 통계를 기록하였다.
- 주로 UN 회원국들이지만 홍콩과 같은 기타 독립 국가들도 포함되어 있다.

• 데이터는 어떻게 이루어져 있는가?

- 총 213개의 관측치가 존재하며, 변수는 7개이다.
- 각각의 변수 명은 region, group, fertility , ppgdp, lifeExpF, pctUrban, infantMortality 이다.

00 분석에 앞서..

설명변수와 반응변수

infantMortality를 반응 변수로 선택하고,
region, group, fertility, ppgdp, lifeExpF, pctUrban 을 설명 변수로 선택

• 데이터 분리 (Train Data + Test Data)

- 전체 213개의 관측치 중 반응변수가 결측값인 자료 6개, 설명변수가 결측값인 자료 14개 제거
- 193(=213−6−14)개의 관측치 중 13개를 Test 데이터로 분리, 180개의 Train 데이터로 분석 진행

• 동일한 데이터를 위한 Seed 고정

- Seed 넘버 : 2014
- Seed에 포함된 ID : 56, 33, 120, 59, 104, 16, 171, 112, 29, 115, 10, 110

Contents_

01.

데이터 소개

Data Explanation

01 데이터 소개

변수 설명

변수명	설명	비고
region	• 지역을 나타내는 변수	범주형
group	• OECD 가입 국가 여부	범주형
fertility	• 출산율(여성당 어린이수)	연속형
ppgdp	• 1인당 국내 총생산 달러(GDP)	연속형
lifeExpF	• 여성의 평균 수명	연속형
pctUrban	• 도시의 비율.	연속형
infantMortality	• 출생 1,000명당 1세 이하의 영유아 사망률	연속형

01 데이터 소개

데이터 확인 (총 213개의 관측치)

	region	group	fertility	ppgdp	lifeExpF	pctUrban	infantMortality
	<fct>	*<fct>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*
Afghanistan	Asia	other	5.97	499.0	49.5	23	124.54
Albania	Europe	other	1.52	3677.2	80.4	53	16.56
Algeria	Africa	africa	2.14	4473.0	75	67	21.46
American Samoa	NA	NA	NA	NA	NA	NA	11.29
Angola	Africa	africa	5.14	4321.9	53.2	59	96.19
Anguilla	Caribbean	other	2.00	13750.1	81.1	100	NA
Argentina	Latin Amer	other	2.17	9162.1	79.9	93	12.34
Armenia	Asia	other	1.74	3030.7	77.3	64	24.27
Aruba	Caribbean	other	1.67	22851.5	77.8	47	14.69
Australia	Oceania	oecd	1.95	57118.9	84.3	89	4.46
…	…	…	…	…	…	…	…

01 데이터 소개

데이터 요약

region	
Africa	53
Aisa	50
Europe	39
Latin America	20
Caribbean	17
(Other)	20
NA's	14

Group	
oecd	31
other	115
africa	53
NA's	14

	fertility	ppgdp	lifeExpF	pctUrban	infantMortality
Min.	1.134	114.8	48.11	11.00	1.916
1st Qu.	1.754	1283.0	65.66	39.00	7.019
Median	2.262	4684.5	75.89	59.00	19.007
Mean	2.761	13012.0	72.29	57.93	29.440
3rd Qu.	3.545	15520.5	79.58	75.00	44.477
Max.	6.925	105095.4	87.12	100.00	124.535
NA's	14	14	14	14	6

01 데이터 소개

• 반응변수 및 설명변수 설정

- infantMortality(영유아 사망률) 을 반응변수로 선택.
- region(지역), group(OECD), fertility(출산율), ppgdp(GDP), lifeExpF(여성평균수명), pctUrban(도시비율) 을 설명변수로 선택.

• 결측값 처리

- 반응변수 결측(6) : Anguilla, Bermuda, Cayman Islands, Dominica, Greenland, Seychelles
- 설명변수 결측(14) : American Samoa, Channel Islands, French Guiana, Guadeloupe, Guam, Martinique, Mayotte, Niue, Northern Mariana Islands, Reunion, Tokelau, United Sates Virgin Islands, Wallis and Funtuna Islands, Westren Sahara

▶총 20개의 결측값 관측자료를 제거하고 분석 진행.

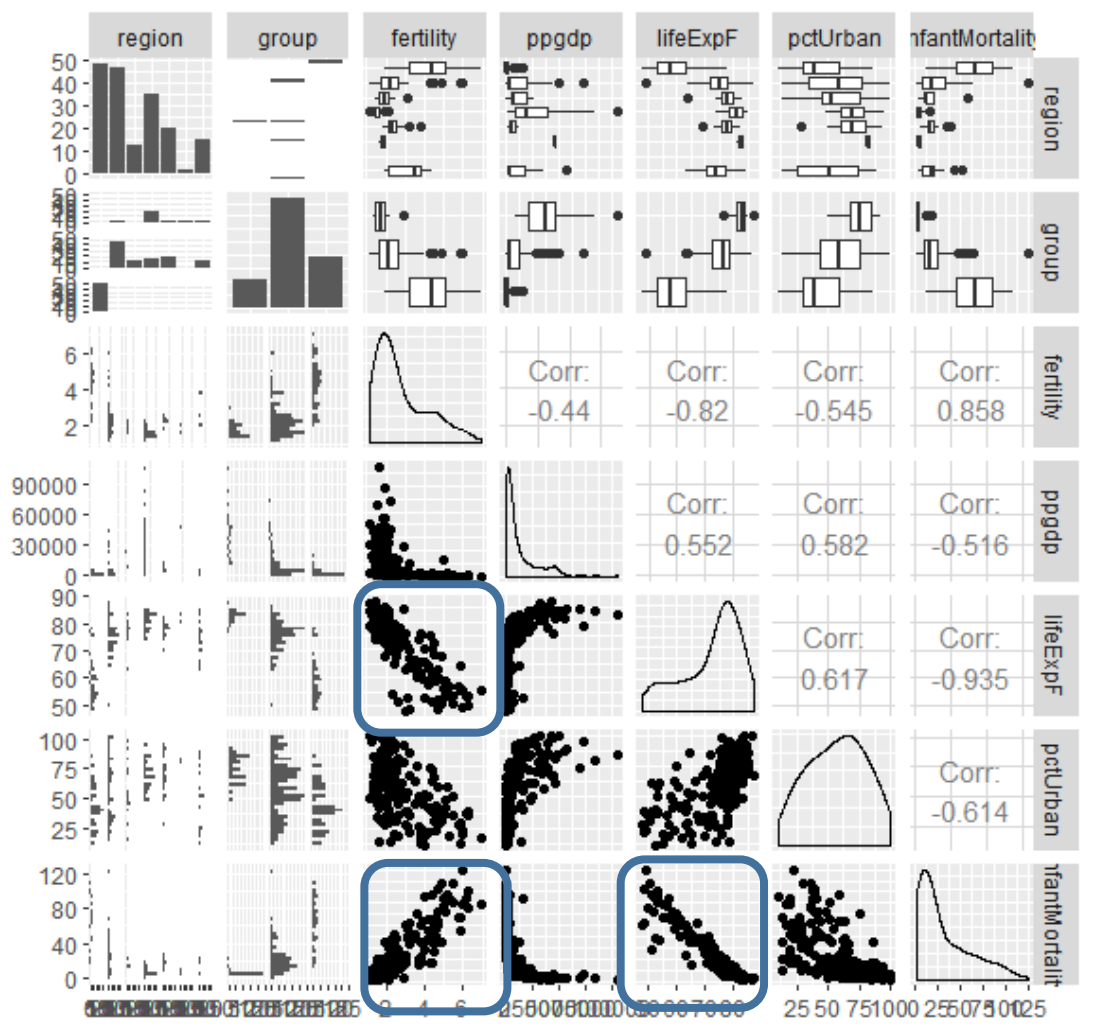
▶180개의 Train Data와 13개의 Test Data로 분리하여 분석 진행.

02.

자료 탐색 (EDA)

Exploratory Data Analysis

02 자료 탐색 (EDA)

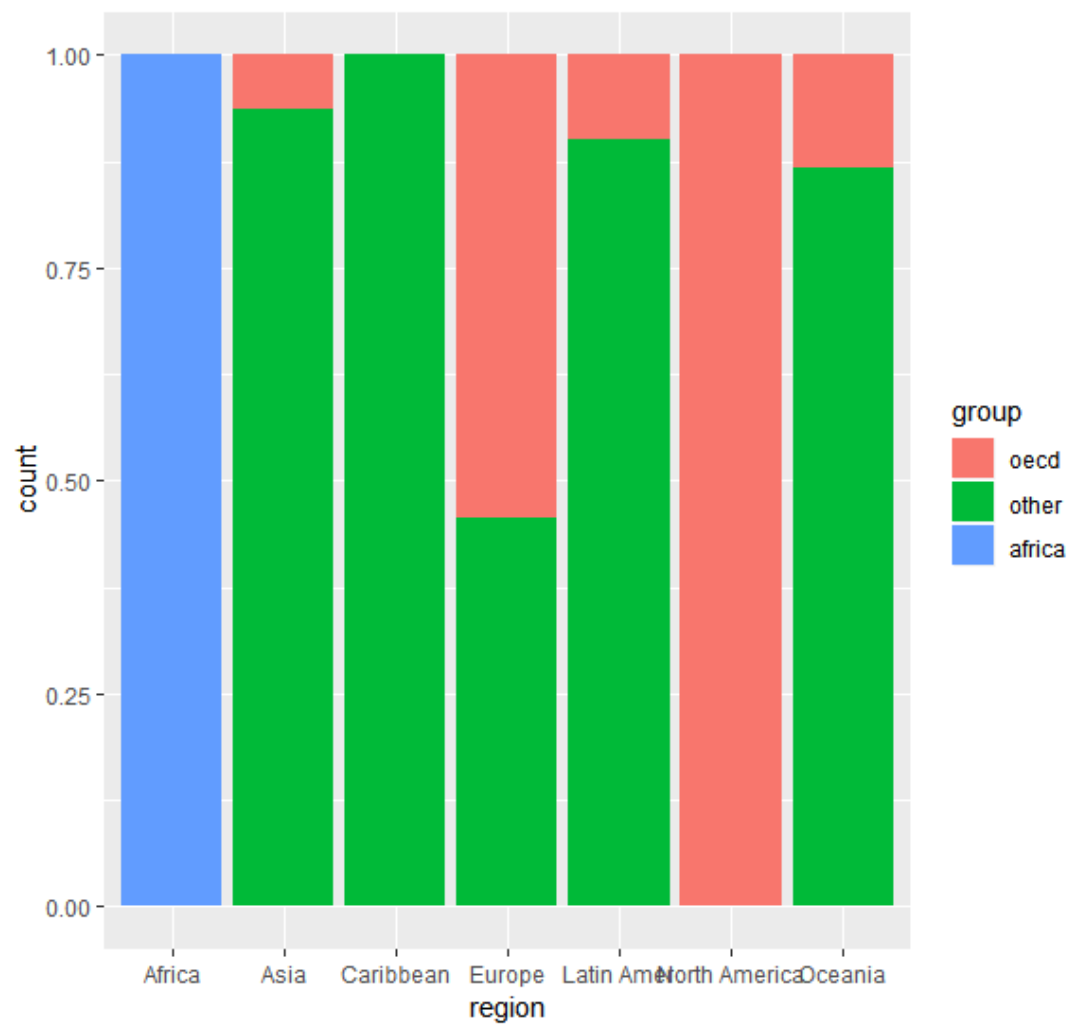


산점도(ggpairs)

상관계수	fertillity	ppgdp	lifeExpF	pctUrban
infantMortality	0.858	−0.516	−0.935	−0.614

- fertility와 infantMortality는 양의 선형 관계를 보인다.
- llifeExpF와 infantMortality 는 음의 선형 관계를 보인다.
- fertility 와 llifeExpF 는 음의 선형 관계를 보인다.
- region 과 group 은 각 범주별 빈도수의 편차가 뚜렷하다.

02 자료 탐색 (EDA)



지역과 OECD 가입 여부의 관계

- Europe 의 56%가 OECD 회원국
- Aisa, Latin America, Oceania 의 일부분만 OECD 회원국
- North America 는 모두 OECD 회원국

Pearson’s Chi-squared test (모의실험 계산)

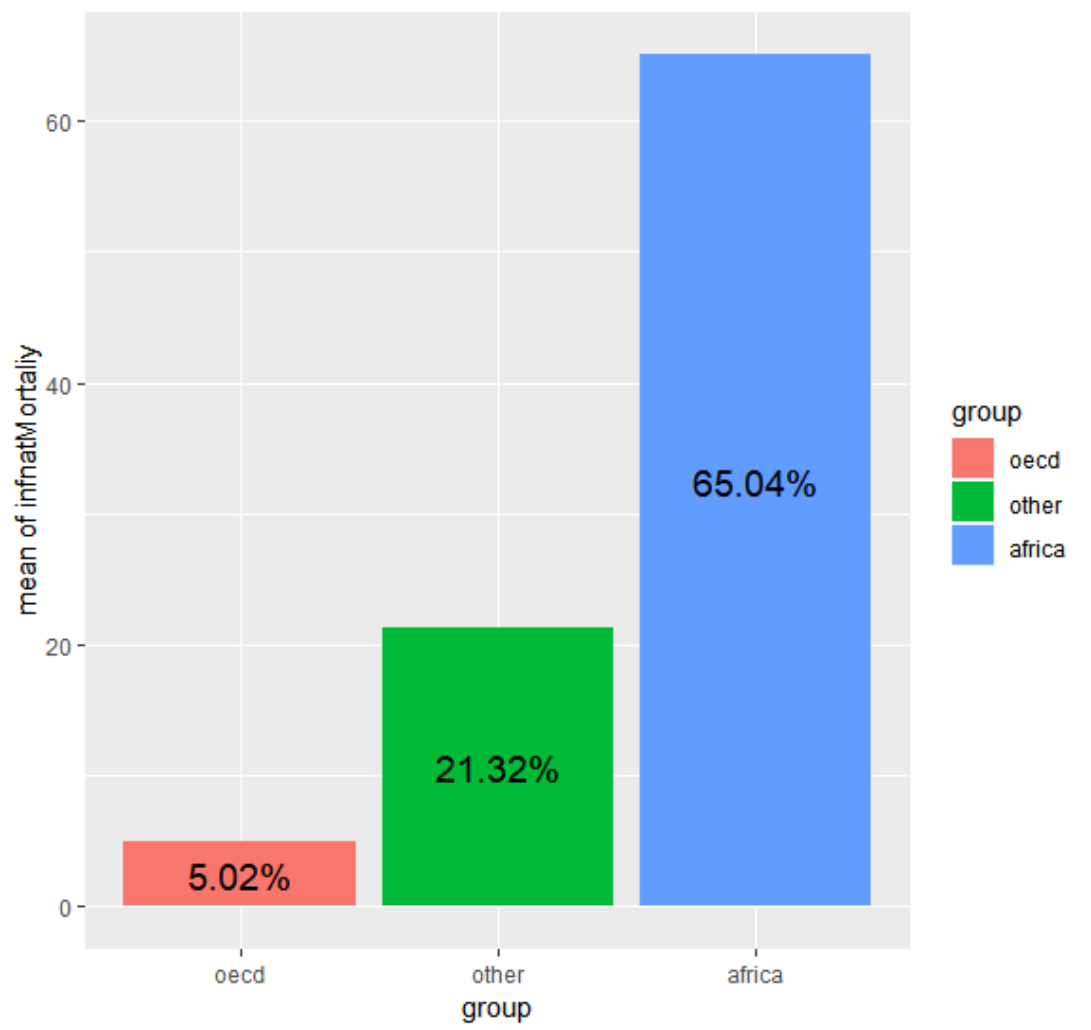
검정통계량	자유도	P-값
237.39	12	2.2×10^{-16}

▶ “두 변수가 서로 독립이다.” 라는 귀무가설을 기각한다.

▶ 두 변수의 연관성이 높아 다중공선성의 문제가 될 수 있다.

▶ 범주가 더 많은 region 변수를 제거 (회귀계수의 개수가 적음.)

02 자료 탐색 (EDA)



OECD 가입 여부와 영유아 사망률의 관계

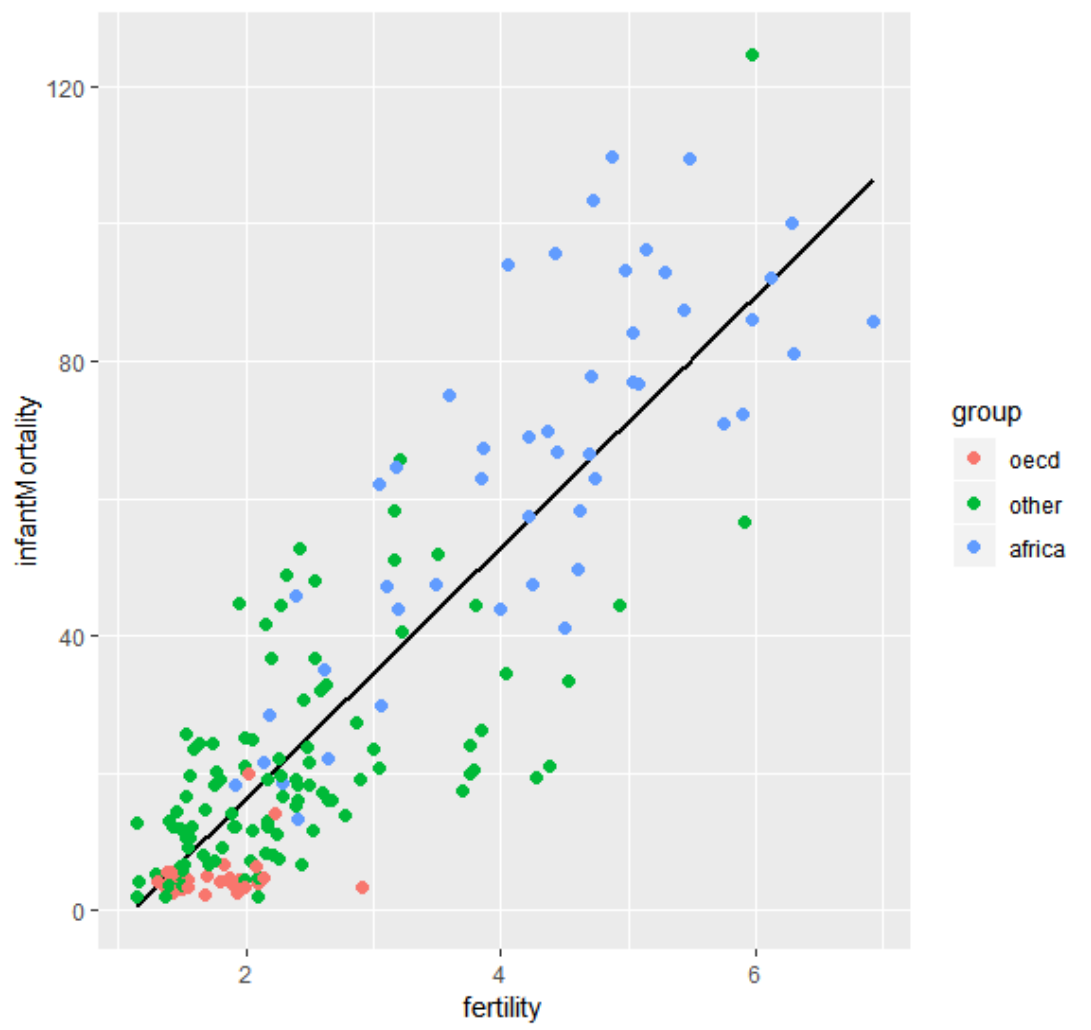
	OECD	other	Africa
국가 수	28	104	48
비율	15.5%	57.7%	26.8%

- OECD 가입 국가의 영유아 사망률의 평균 : 5.02%
- 기타 국가의 영유아 사망률 평균 : 21.32%
- Africa 국가의 영유아 사망률 평균 : 65.04%
- 전체 영유아 사망률 평균 : 30.4%

▶ 아프리카 국가들의 영유아 사망률이 현저하게 높다.

▶ OECD 가입 여부에 따라 평균 영유아 사망률의 차이가 크다.

02 자료 탐색 (EDA)



출산율과 영유아 사망률의 관계

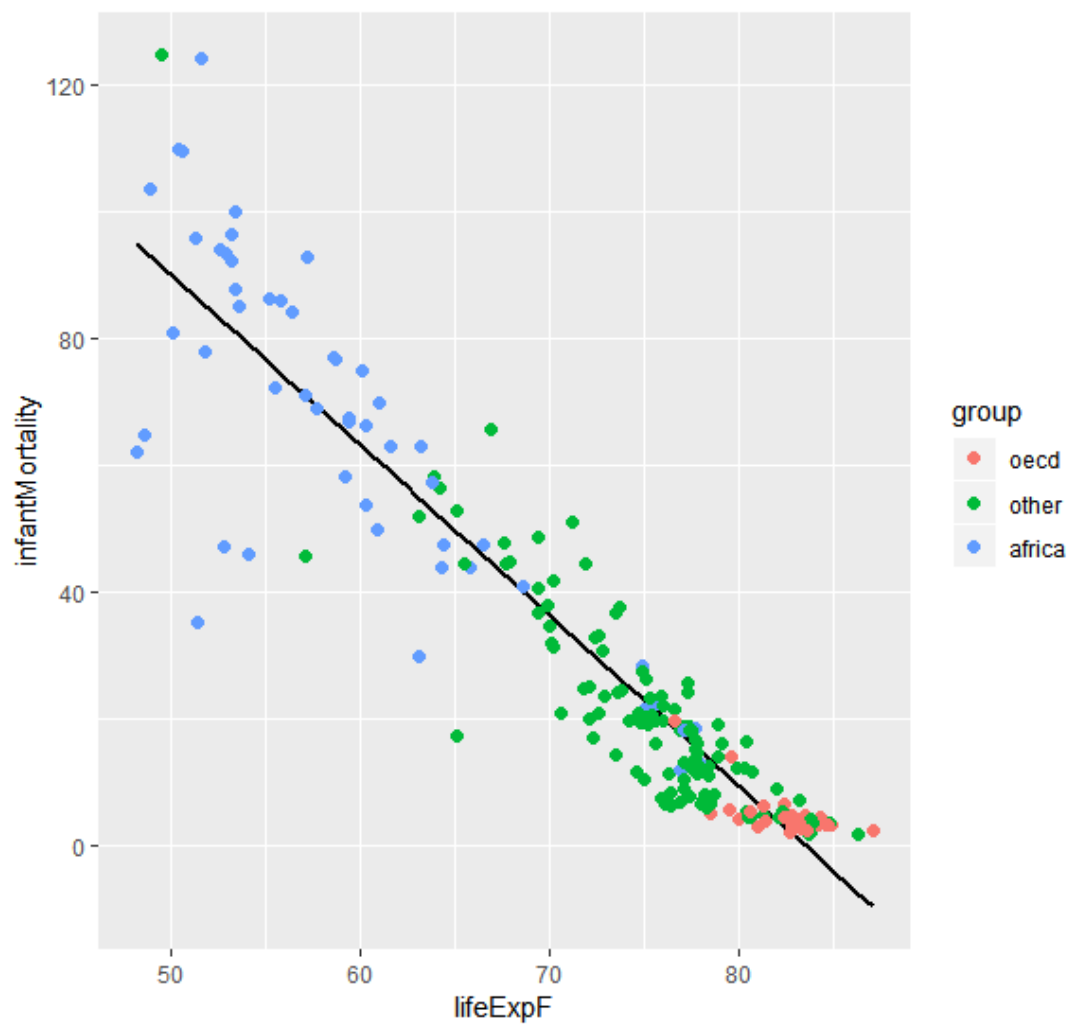
- OECD 가입 국가의 평균 출산율: 1.76명
- 기타 국가의 평균 출산율 : 2.35명
- Africa 국가의 평균 출산율: 4.27명
- 전체 평균 출산율 : 2.77명

▶ OECD 가입 국가는 출산율, 영유아의 사망률 모두 낮았다.

▶ Africa 국가는 출산율, 영유아의 사망률 모두 높았다.

▶ 출산율과 영유아 사망률의 관계는 양의 선형 관계를 보인다.

02 자료 탐색 (EDA)



여성의 평균 수명과 영유아 사망률의 관계

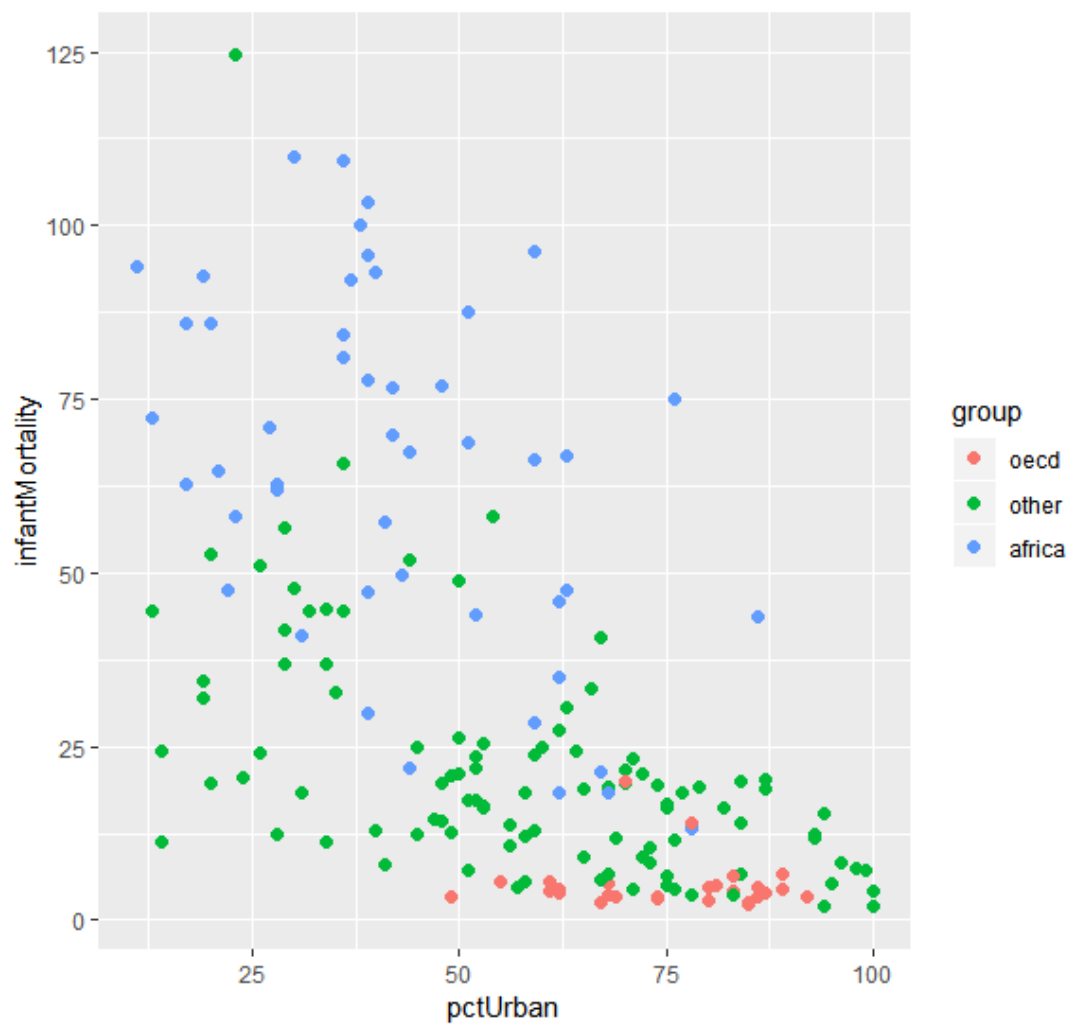
- OECD 가입 국가의 평균 여성 수명: 82.5년
- 기타 국가의 평균 여성 수명 : 75.4년
- Africa 국가의 평균 여성 수명 : 59.3년
- 전체 평균 여성 수명 : 72.2년

▶ OECD 가입 국가는 기대 수명이 전체 평균보다 높다.

▶ Africa 국가는 여성의 기대 수명이 전체 평균보다 낮았다.

▶ 여성의 수명과 영유아 사망률의 관계는 음의 선형 관계를 보인다.

02 자료 탐색 (EDA)



도시 비율과 영유아 사망률의 관계

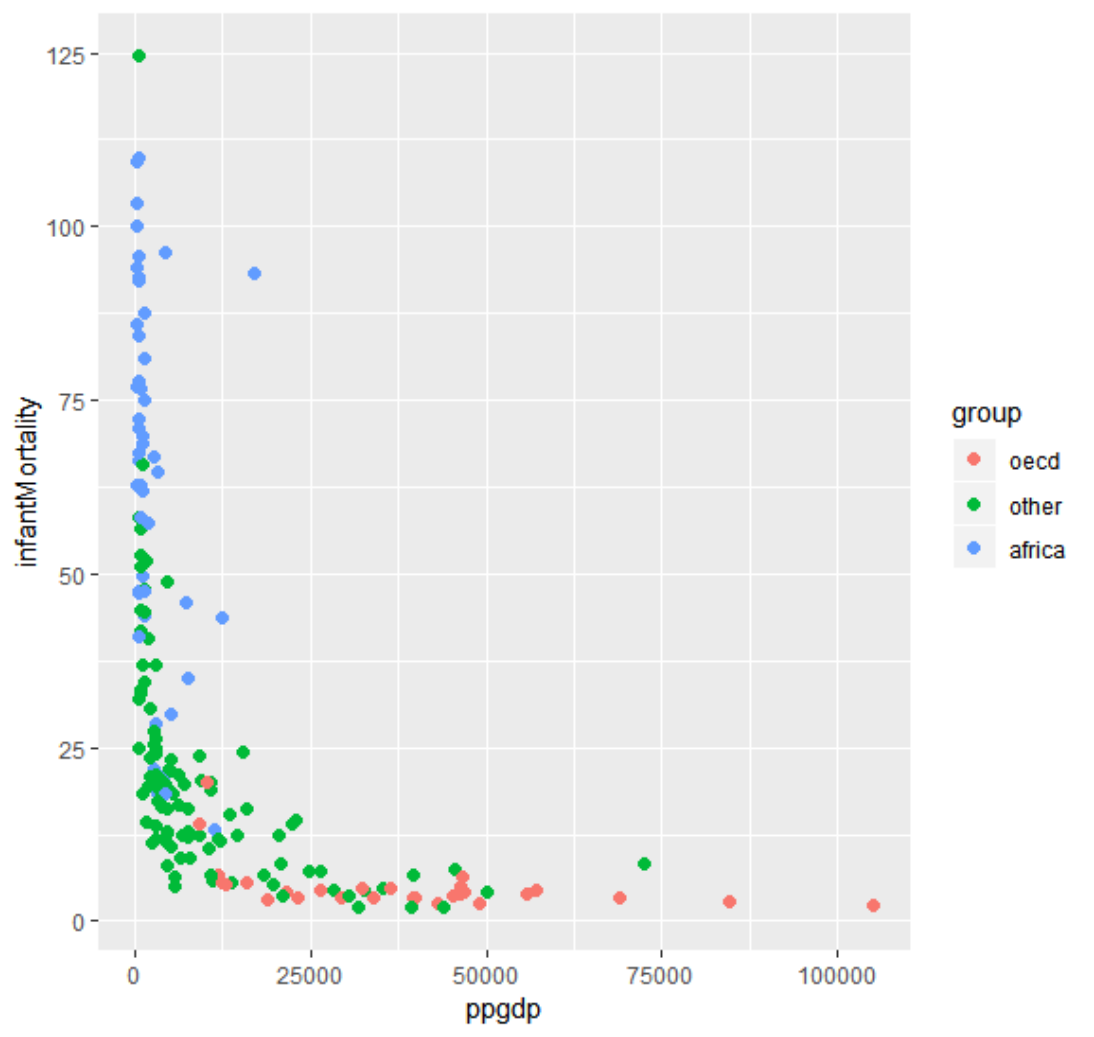
- OECD 가입 국가의 평균 도시 비율 : 74.9%
- 기타 국가의 평균 도시 비율 : 58.9%
- Africa 국가의 평균 도시 비율 : 42.6%
- 전체 평균 도시 비율 :57.1%

▶ OECD 가입 국가의 도시 비율이 평균에 비해 약 17% 높았다.

▶ group에 따라 층이 나눠졌다.

▶pctUrban은 회귀 분석 진행 시 중요하지 않을 것으로 예상된다.

02 자료 탐색 (EDA)



GDP와 영유아 사망률의 관계

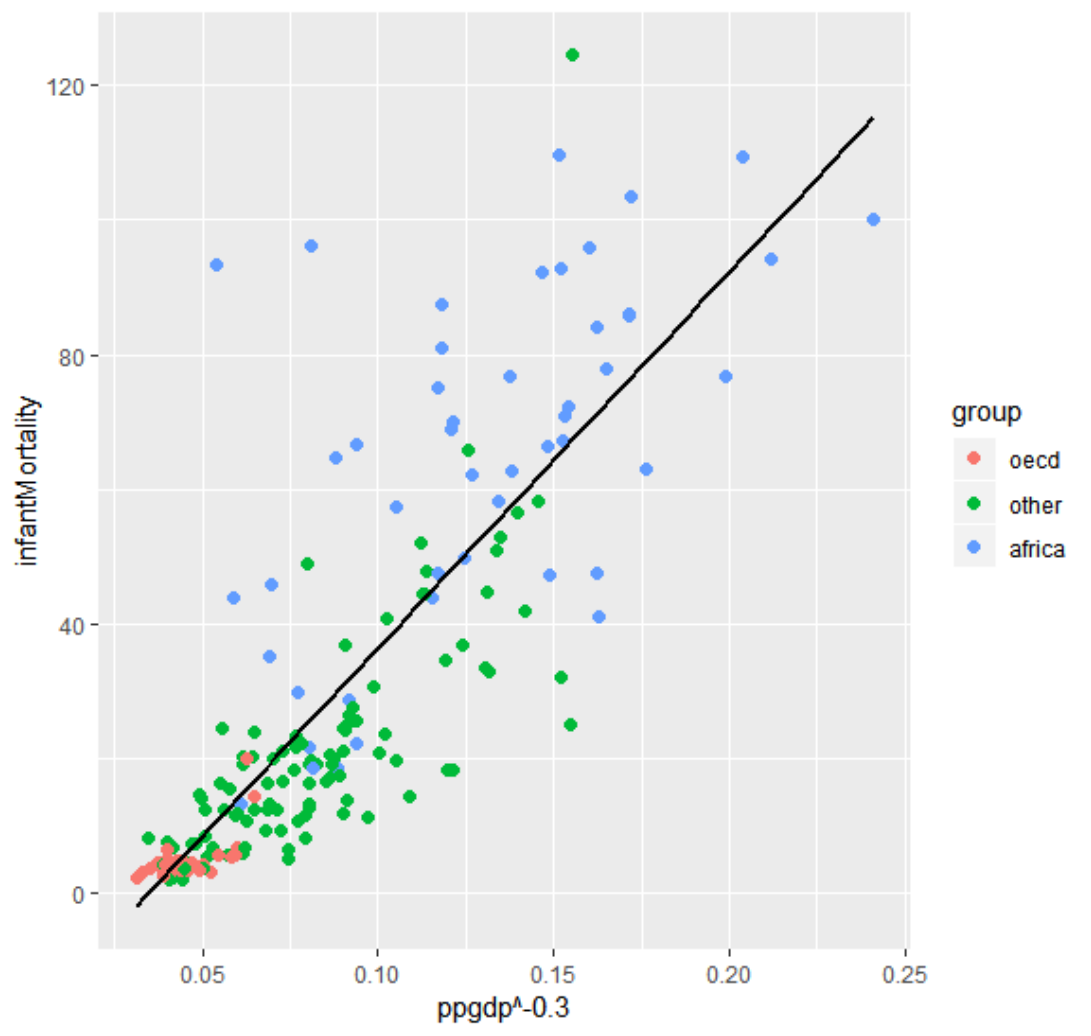
- OECD 가입 국가의 평균 GDP : $ 38,148
- 기타 국가의 평균 GDP : $ 10,083
- Africa 국가의 평균 GDP : $ 2,324
- 전체 평균 GDP : $ 12,379

▶ group에 따라 데이터의 scale의 편차가 큰 편이다.

▶ 관계를 명확하게 하기 위해 적절한 변수 변환이 필요해 보인다.

▶ boxTidwell에 의한 설명변수 변환 람다 추정 : $\hat{\lambda} = -0.3$

02 자료 탐색 (EDA)



$GDP^{-0.3}$ 과 영유아 사망률의 관계

- OECD 가입 국가의 평균 $GDP^{-0.3}$: 0.0455
- 기타 국가의 평균 $GDP^{-0.3}$: 0.0824
- Africa 국가의 평균 $GDP^{-0.3}$: 0.1290
- 전체 평균 $GDP^{-0.3}$: 0.0891

▶ group에 따라 $GDP^{-0.3}$의 차이가 존재한다.

▶ $GDP^{-0.3}$ 가 증가할수록 영유아의 사망률이 증가한다.

03.

변수 선택

Variable Select

03 변수 선택

후진소거법에 의한 잠정모형 추정

- 모든 변수가 포함된 모형에서 비유의적인 변수를 차례로 제거

첫 번째 변수 제거

	group	fertiltiy	lifeExpF	$ppgdp^{-0.3}$	pctUrban
Sum of Sq	295.1	1818.6	11034.2	1734.4	49.4
P-Value	0.1399	1.738×10^{-6}	2.2×10^{-16}	2.917×10^{-6}	0.4155
제거 변수					제거

두 번째 변수 제거

	group	fertiltiy	lifeExpF	$ppgdp^{-0.3}$
Sum of Sq	291.8	1849.1	11265.3	1861.2
P-Value	0.1424	1.405×10^{-6}	2.2×10^{-16}	1.304×10^{-6}
제거 변수	제거			

03 변수 선택

후진소거법에 의한 잠정모형 추정

- 모든 변수가 포함된 모형에서 비유의적인 변수를 차례로 제거

세 번째 변수 제거

	fertiltiy	lifeExpF	$ppgdp^{-0.3}$
Sum of Sq	2121.1	14946.8	1635.7
P-Value	3.072×10^{-7}	2.2×10^{-16}	5.83×10^{-6}
제거 변수	제거할 변수 없음.		

잠정모형

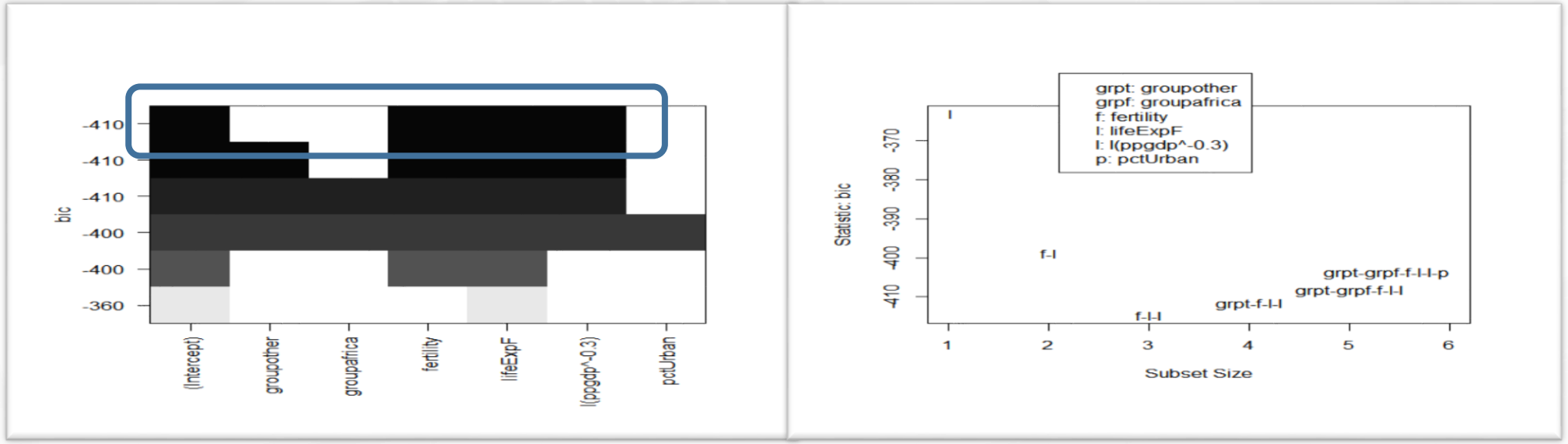
- fertility(출산율), lifeExpF(여성평균수명), $ppgdp^{-0.3}$ (GDP)가 포함된 모형

- $Y_{infantMortality} = \beta_0 + \beta_1 X_{fertility} + \beta_2 X_{lifeExpF} + \beta_3 X_{ppgdp\hat{}-0.3}$

03 변수 선택

BIC를 기준으로 변수 선택

- 모든 가능한 회귀에서 BIC 값을 최소로 하는 모델 선택



잠정모형

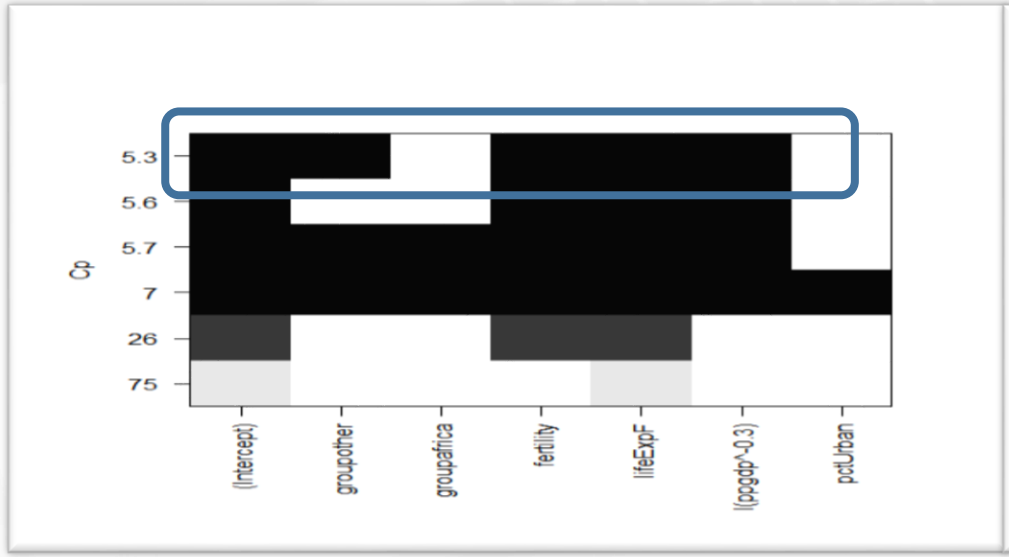
- fertility(출산율), lifeExpF(여성평균수명), $ppgdp^{-0.3}$(GDP)가 포함된 모형

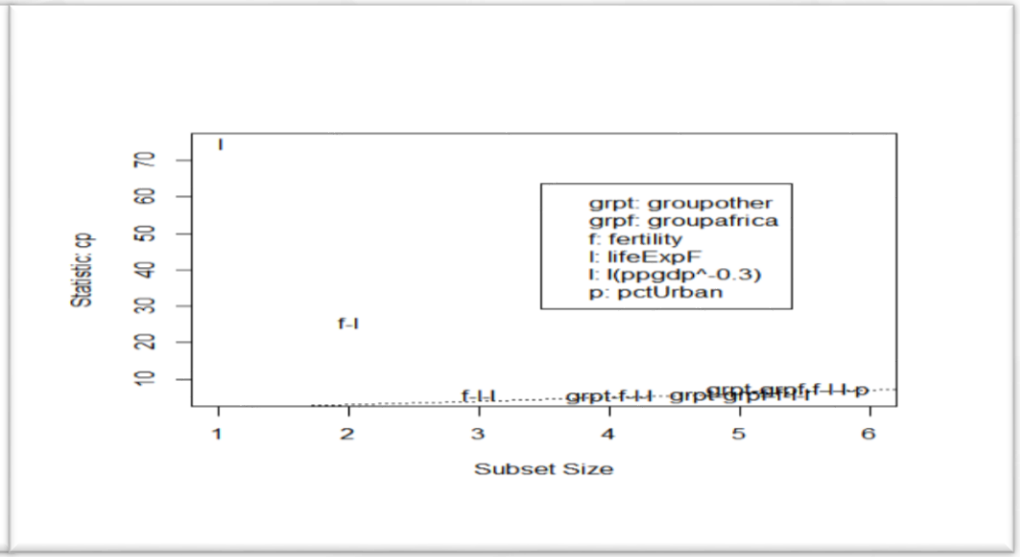
- $Y_{infantMortality} = \beta_0 + \beta_1 X_{fertility} + \beta_2 X_{lifeExpF} + \beta_3 X_{ppgdp\text{^}-0.3}$

03 변수 선택

C_p통계량을 기준으로 변수 선택

- $C_p \approx p$ 에 가장 근접한 모델 선택





잠정모형

- group(OECD), fertility(출산율), lifeExpF(여성평균수명), $ppgdp^{-0.3}$(GDP)가 포함된 모형

- $Y_{\text{infantMortality}} = \beta_0 + \beta_1 D_{groupother} + \beta_2 D_{groupafrica} + \beta_3 X_{\text{fertility}} + \beta_4 X_{\text{lifeExpF}} + \beta_5 X_{\text{ppgdp^}-0.3}$

03 변수 선택

region과 ppgdp를 그대로 사용한다면?

- region, group, ppgdp, lifeExpF, pctUrban을 설명변수로 사용했을 때의 변수 선택의 결과

변수의 유의성 기준의 후진 소거법

- ppgdp가 첫 번째로 제거, group이 두 번째로 제거
- $Y_{infantMortality} = \beta_0 + \beta_1 D_{regionAsia} + \cdots + \beta_6 D_{regionOceania} + \beta_7 X_{fertility} + \beta_8 X_{lifeExpF} + \beta_9 X_{pctUrban}$

모형 설정 기준에 의한 변수 선택

- BIC 기준에 의한 선택 : $Y_{infantMortality} = \beta_0 + \beta_1 D_{gorupother} + \beta_2 D_{gorupafrica} + \beta_3 X_{ppgdp} + \beta_4 X_{pctUrban}$
- C_p 통계량 기준에 의한 선택 : $Y_{infantMortality} = \beta_0 + \beta_1 D_{gorupother} + \beta_2 D_{gorupafrica} + \beta_3 X_{ppgdp} + \beta_4 X_{pctUrban}$

03 변수 선택

잠정 모형

- 잠정 모형 1, 2 : region을 제거하고, $ppgdp^{-0.3}$를 사용했을 때의 모형
- 잠정 모형 3, 4 : region와 ppgdp를 그대로 사용했을 때의 모형

잠정 모형 1

- $Y_{infantMortality} = \beta_0 + \beta_1 X_{fertility} + \beta_2 X_{lifeExpF} + \beta_3 X_{ppgdp\text{^}-0.3}$

잠정 모형 2

- $Y_{infantMortality} = \beta_0 + \beta_1 D_{groupother} + \beta_2 D_{groupafrica} + \beta_3 X_{fertility} + \beta_4 X_{lifeExpF} + \beta_5 X_{ppgdp\text{^}-0.3}$

잠정 모형 3

- $Y_{infantMortality} = \beta_0 + \beta_1 D_{regionAsia} + \cdots + \beta_6 D_{regionOceania} + \beta_7 X_{fertility} + \beta_8 X_{lifeExpF} + \beta_9 X_{pctUrban}$

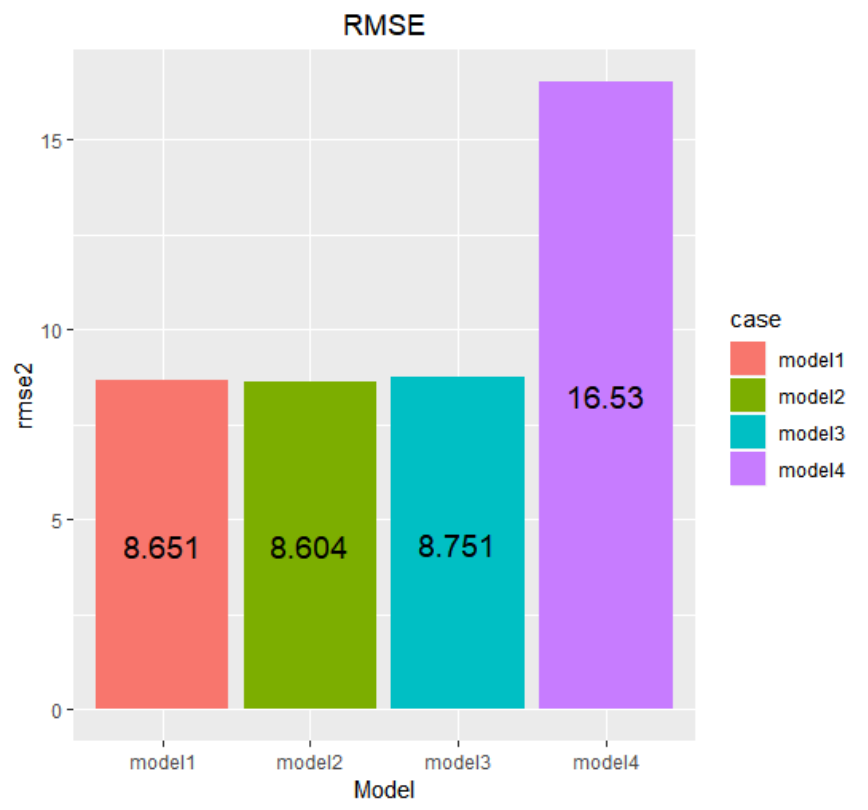
잠정 모형 4

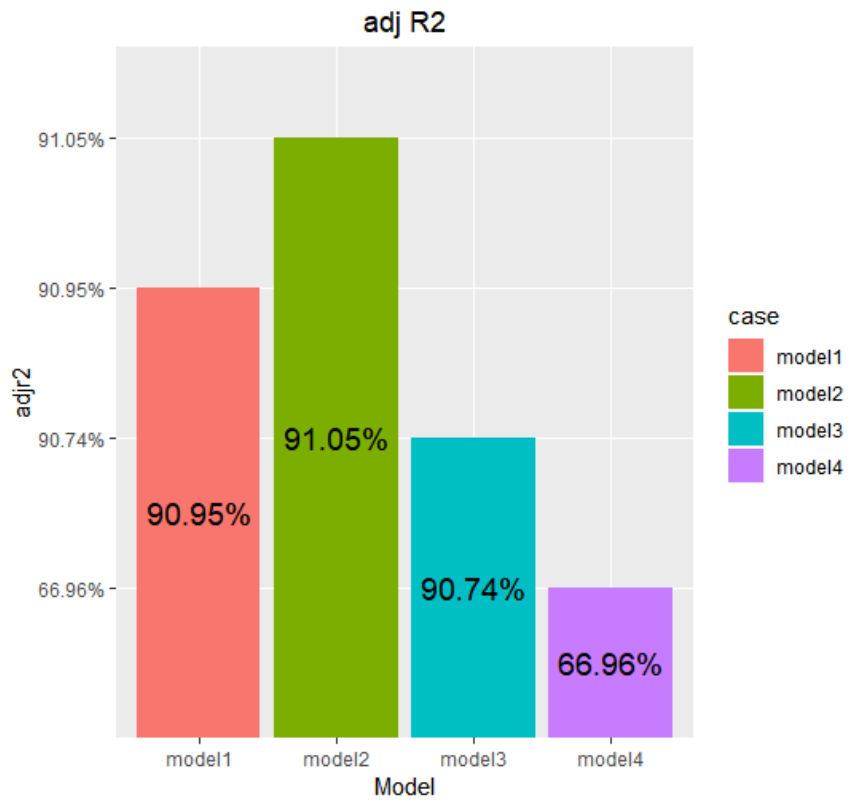
- $Y_{infantMortality} = \beta_0 + \beta_1 D_{groupother} + \beta_2 D_{groupafrica} + \beta_3 X_{ppgdp} + \beta_4 X_{pctUrban}$

03 변수 선택

잠정 모형

- 잠정 모형 별 RMSE, 수정된 결정계수 비교





▶ Model1, 2, 3 은 큰 차이가 없었고, Model 4 가 나머지에 비해 큰 차이를 보였다.

04.

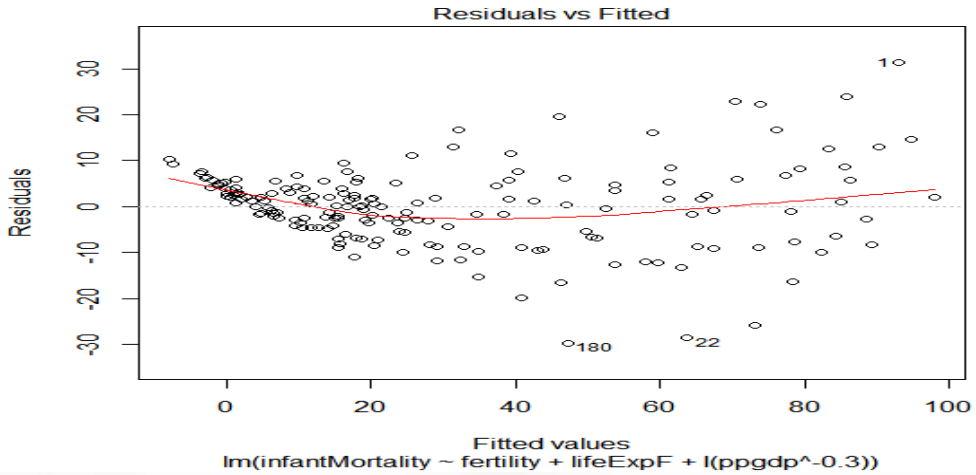
잠정 모형 확인

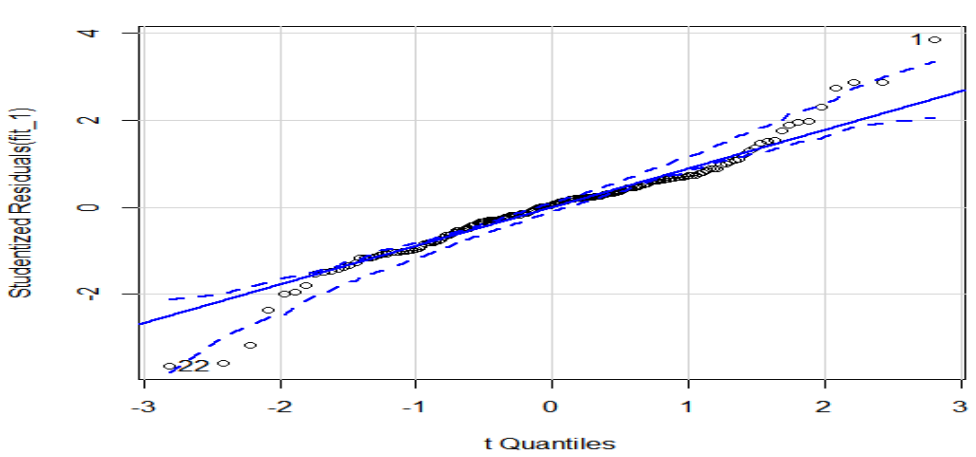
Assumption Check

04 잠정 모형 확인

첫 번째 잠정 모형

- $Y_{infantMortality} = 133.13 + 4.68\,X_{fertility} - 1.75\,X_{lifeExpF} + 120.11\,X_{ppgdp\text{^}-0.3}$
- 유의수준 10% 하에서, Intercept, fertility, lifeExpF, $ppgdp^{-0.3}$ 모든 변수가 유의적





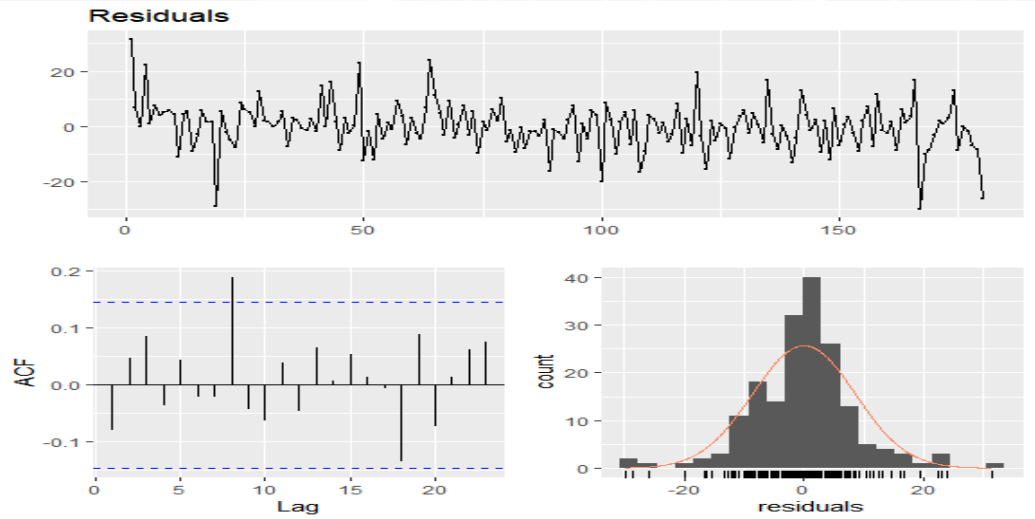
▶ 분산이 점점 증가는 것으로 보인다. → 반응 변수의 변환

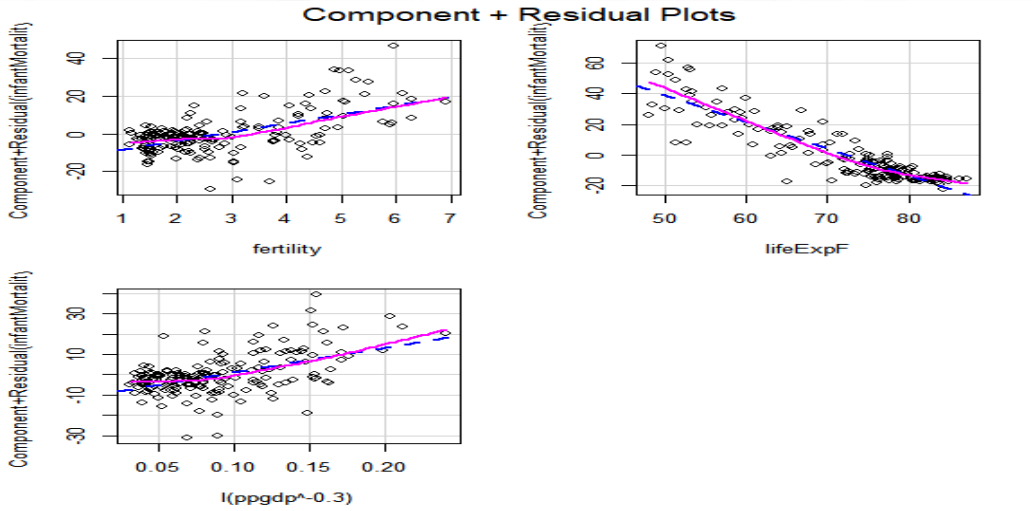
▶ 그래프로 볼 때, 정규성은 만족한 것으로 보인다.

04 잠정 모형 확인

첫 번째 잠정 모형

- $Y_{infantMortality} = 133.13 + 4.68\,X_{fertility} - 1.75\,X_{lifeExpF} + 120.11\,X_{ppgdp\text{^}-0.3}$
- 유의수준 10% 하에서, Intercept, fertility, lifeExpF, $ppgdp^{-0.3}$ 모든 변수가 유의적





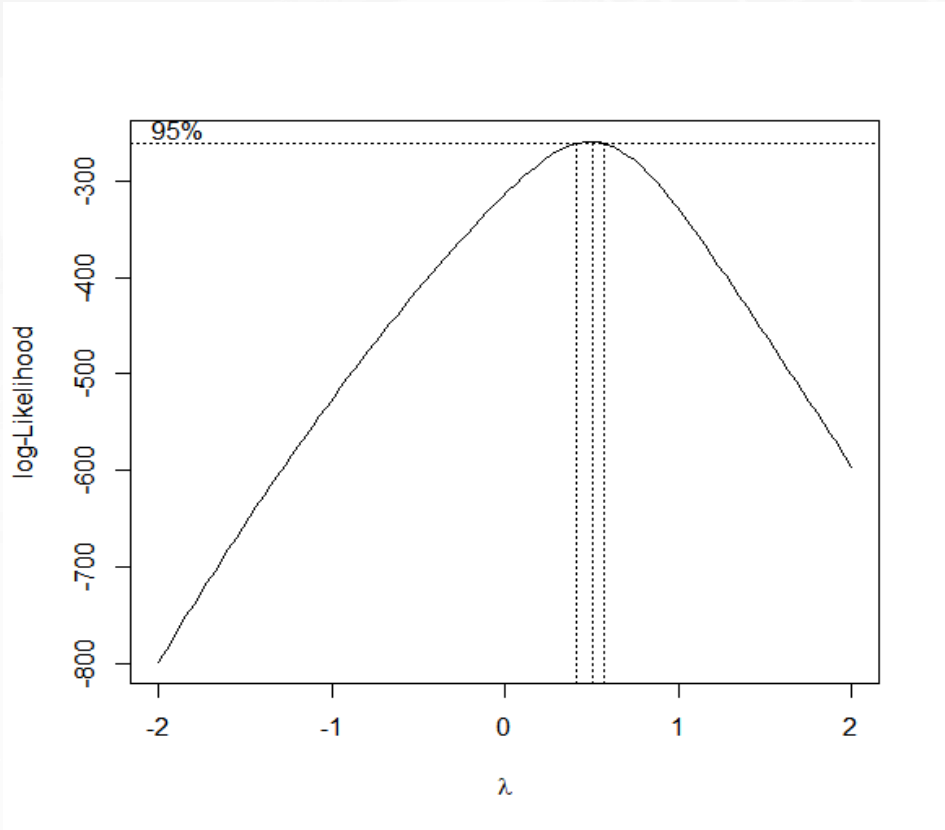
▶ Breusch–Godfrey test 결과 p–value가 0.301로 독립성은 만족하는 것으로 보인다.

▶ 반응 변수와 설명 변수간의 선형 관계 역시 만족하는 것으로 보여진다.

04 잠정 모형 확인

첫 번째 잠정 모형의 반응 변수 변환

• 동일 분산 가정을 만족시키기 위해 반응 변수의 변환이 필요하다.



• 추정된 $\hat{\lambda} = 0.5$

• $\hat{\lambda} = 1$ 에 대한 가설 검정 : 귀무가설 기각. 변수 변환 필요

• $\hat{\lambda} = 0$ 에 대한 가설 검정 : 귀무가설 기각.
로그변환이 아닌 추정된 $\hat{\lambda}$ 에 의한 변수 변환

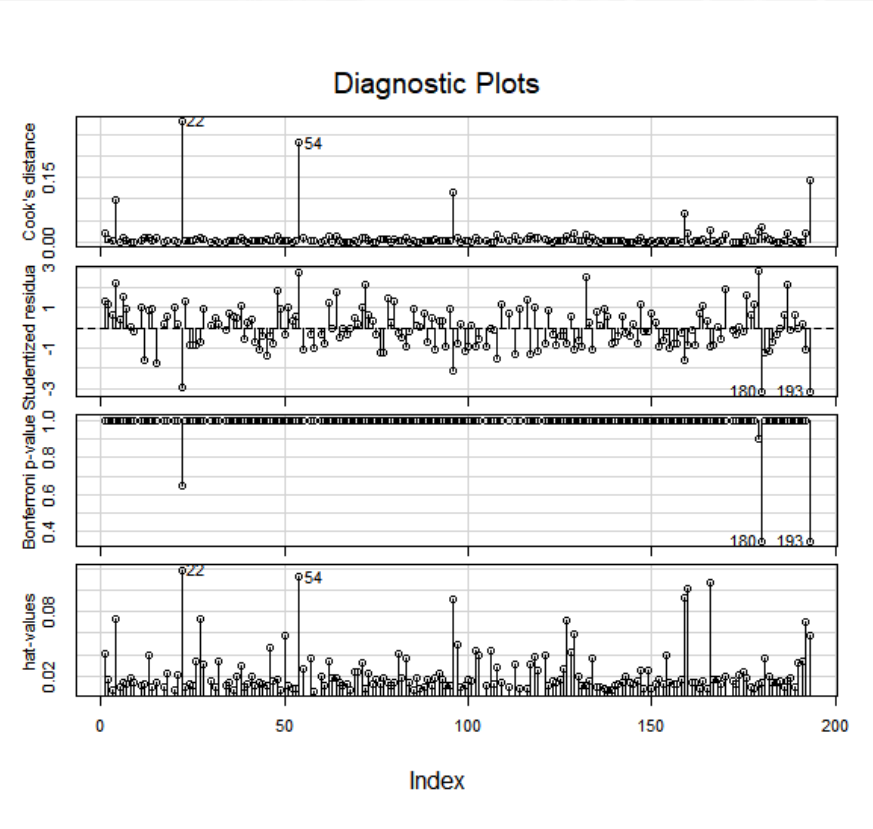
▶ $\hat{\lambda} = 0.5$ 즉, 루트 변환 실시.

▶ $y \rightarrow \sqrt{y}$

04 잠정 모형 확인

첫 번째 잠정 모형의 특이한 관찰값 탐지

- $\sqrt{Y_{infantMortality}} = 133.13 + 4.68\,X_{fertility} - 1.75\,X_{lifeExpF} + 120.11\,X_{ppgdp\text{^}-0.3}$

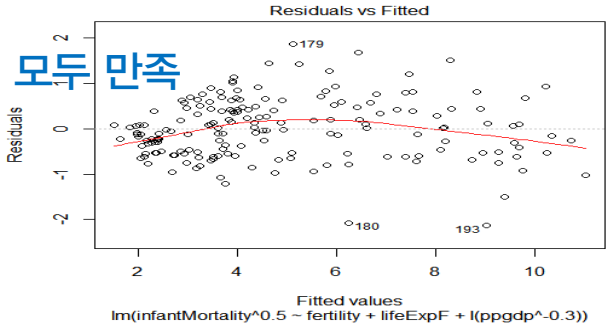


- 특이한 관찰값 후보 : Botswana(22), Equatorial Guinea(54)

- 특이한 관찰값 제거 후 회귀 모형 적합

▶ Intercept, fertility, lifeExpF, $ppgdp^{-0.3}$ 모든 변수가 유의적

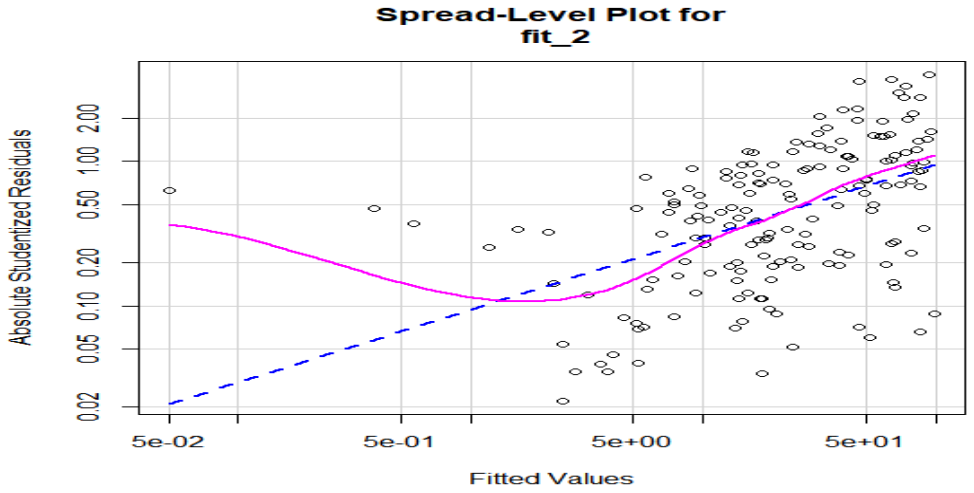
▶ 등분산, 정규성, 독립성, 선형 관계 모두 만족

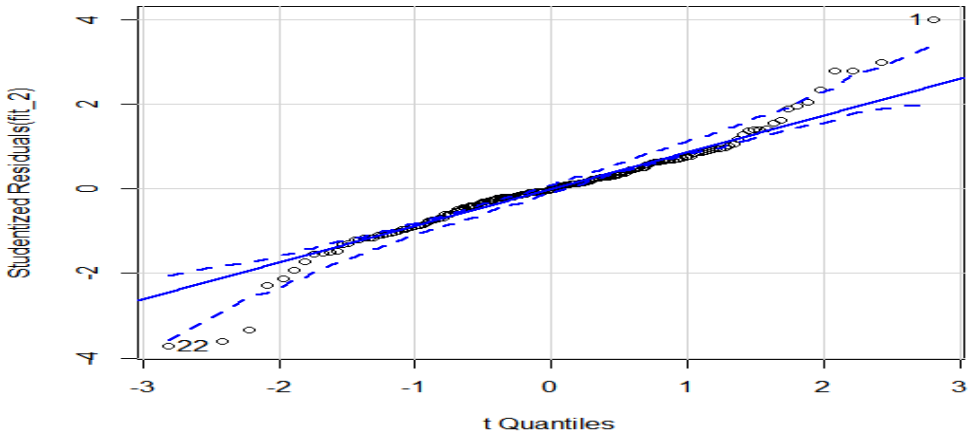


04 잠정 모형 확인

두 번째 잠정 모형

- $Y_{infantMortality} = 140.08 - 4.02\, D_{groupother} - 4.10\, D_{groupafrica} + 4.44\, X_{fertility} - 1.80\, X_{lifeExpF} + 131.65\, X_{ppgdp^{-0.3}}$
- 유의수준 10% 하에서, Intercept, group, fertility, lifeExpF, $ppgdp^{-0.3}$ 모든 변수가 유의적





▶ 분산이 점점 증가는 것으로 보인다. → 반응 변수의 변환

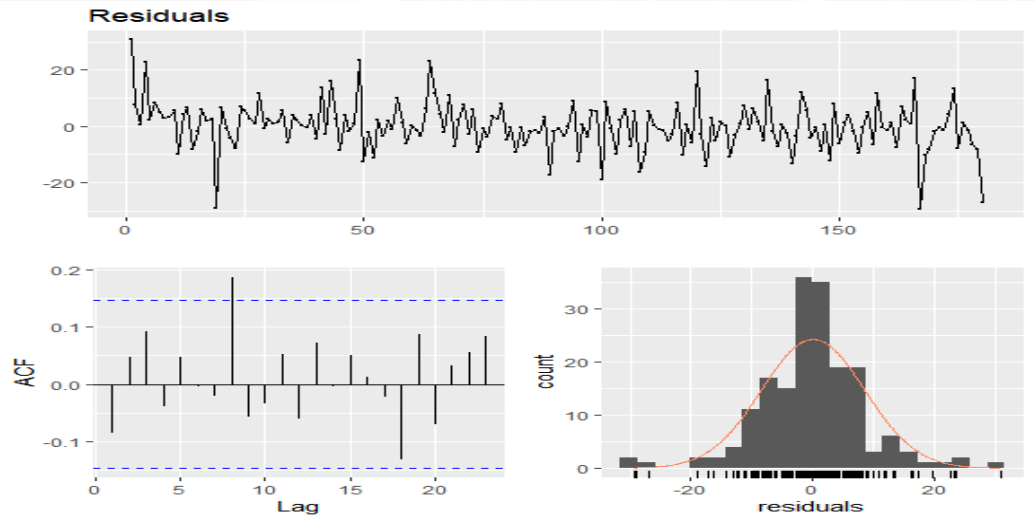
▶ 그래프로 볼 때, 정규성은 만족한 것으로 보인다.

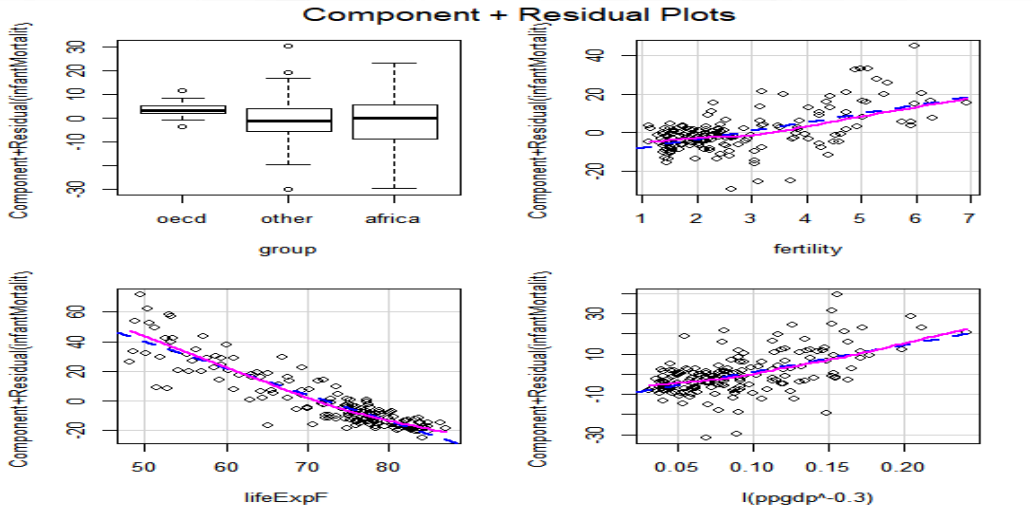
04 잠정 모형 확인

두 번째 잠정 모형

- $Y_{infantMortality} = 140.08 - 4.02\, D_{groupother} - 4.10\, D_{groupafrica} + 4.44\, X_{fertility} - 1.80\, X_{lifeExpF} + 131.65\, X_{ppgdp^{-0.3}}$

- 유의수준 10% 하에서, Intercept, group, fertility, lifeExpF, $ppgdp^{-0.3}$ 모든 변수가 유의적





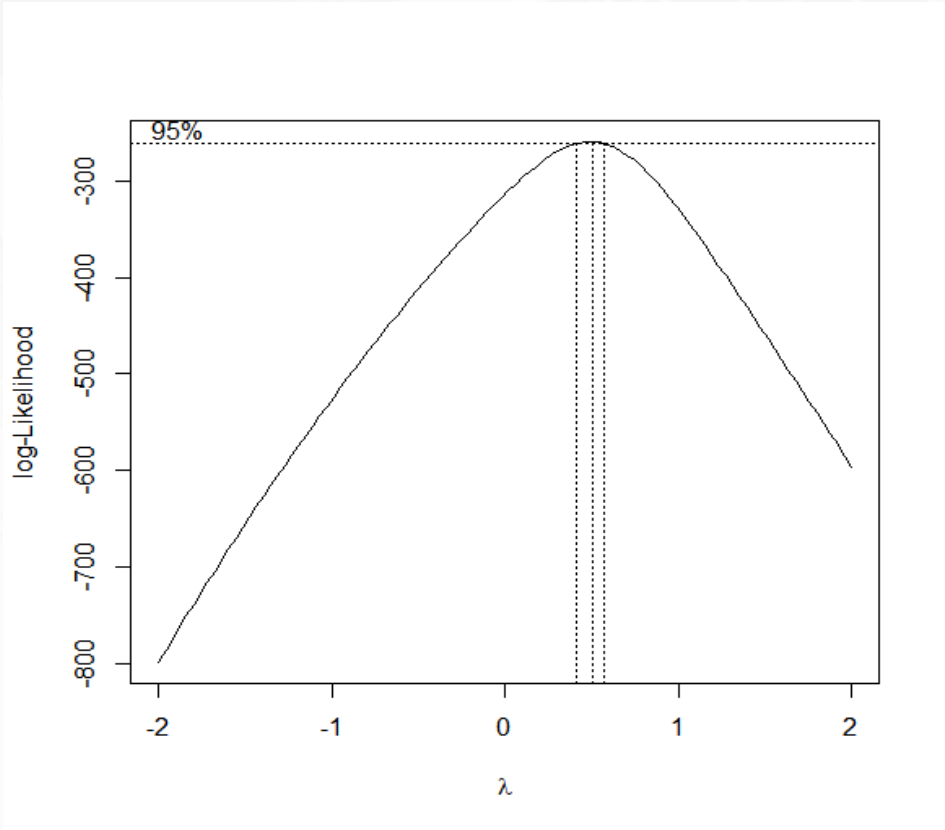
▶ Breusch–Godfrey test 결과 p–value가 0.3461로 독립성은 만족하는 것으로 보인다.

▶ 반응 변수와 설명 변수간의 선형 관계 역시 만족하는 것으로 보여진다.

04 잠정 모형 확인

두 번째 잠정 모형의 반응 변수 변환

• 동일 분산 가정을 만족시키기 위해 반응 변수의 변환이 필요하다.



• 추정된 $\hat{\lambda} = 0.5$

• $\hat{\lambda} = 1$ 에 대한 가설 검정 : 귀무가설 기각. 변수 변환 필요

• $\hat{\lambda} = 0$ 에 대한 가설 검정 : 귀무가설 기각.
로그변환이 아닌 추정된 $\hat{\lambda}$ 에 의한 변수 변환

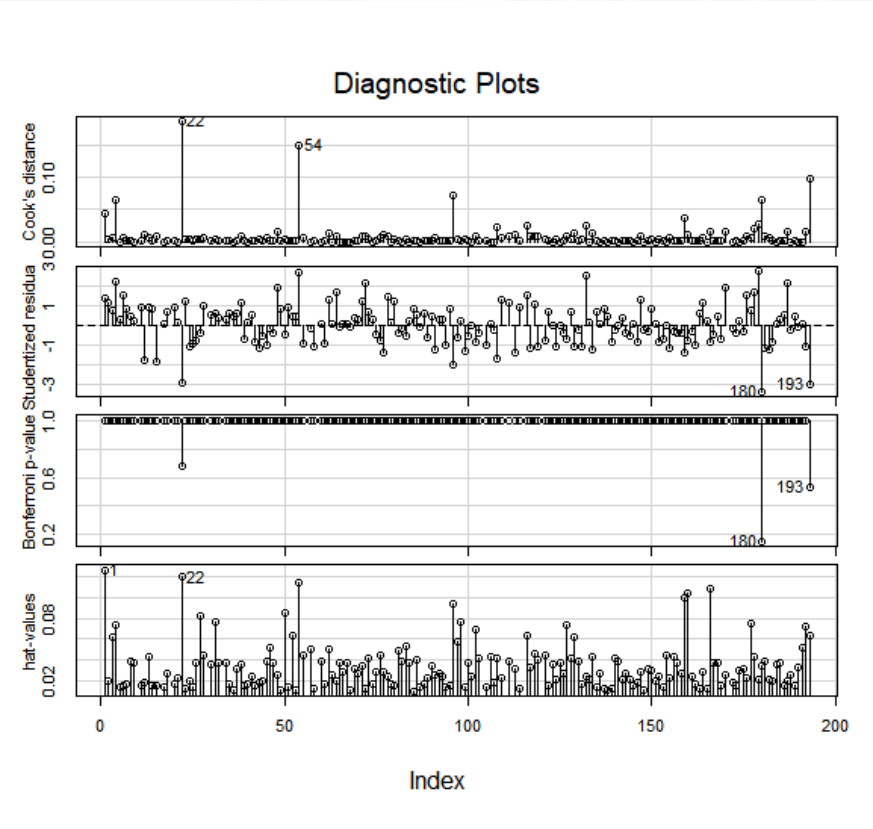
▶ $\hat{\lambda} = 0.5$ 즉, 루트 변환 실시.

▶ $y \rightarrow \sqrt{y}$

04 잠정 모형 확인

두 번째 잠정 모형의 특이한 관찰값 탐지

- $\sqrt{Y_{infantMortality}} = 140.08 - 4.02\, D_{groupother} - 4.10\, D_{groupafrica} + 4.44\, X_{fertility} - 1.80\, X_{lifeExpF} + 131.65\, X_{ppgdp^{-0.3}}$

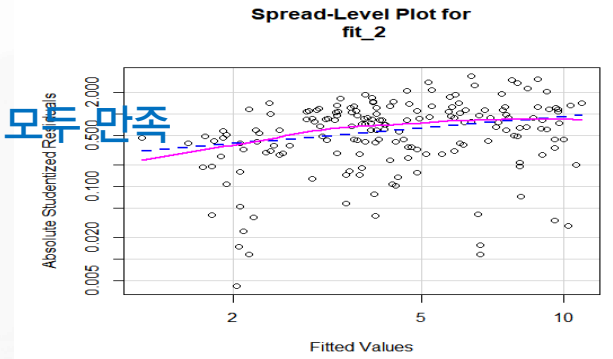


- 특이한 관찰값 후보 : Botswana(22), Tuvalu(180)

- 특이한 관찰값 제거 후 회귀 모형 적합

▶ Intercept, group, fertility, lifeExpF, $ppgdp^{-0.3}$ 모든 변수가 유의적

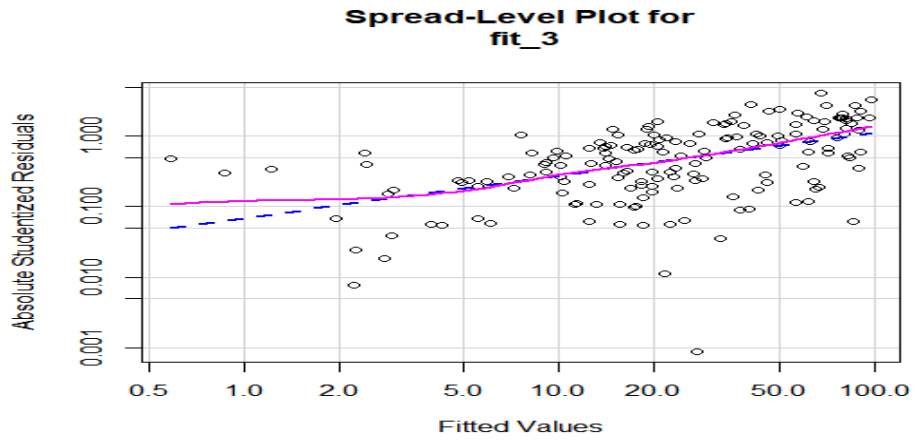
▶ 등분산, 정규성, 독립성, 선형 관계 모두 만족

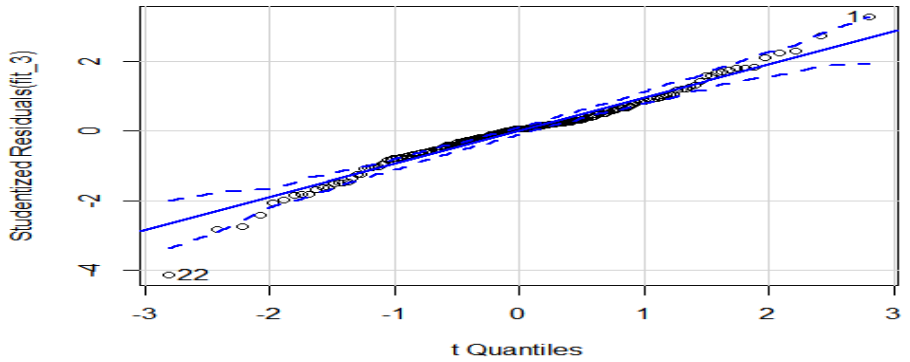


04 잠정 모형 확인

세 번째 잠정 모형

- $Y_{infantMortality} = 152.07 + 2.40\, D_{regionAsia} + \cdots - 9.16\, D_{regionOceania} + 6.56\, X_{fertility} - 1.89\, X_{lifeExpF} - 0.06\, X_{pctUrban}$
- 유의수준 10% 하에서, Intercept, region, fertility, lifeExpF, pctUrban 모든 변수가 유의적





▶ 분산이 증가하거나 감소하는 것처럼 보이지 않는다. 동일 분산 만족

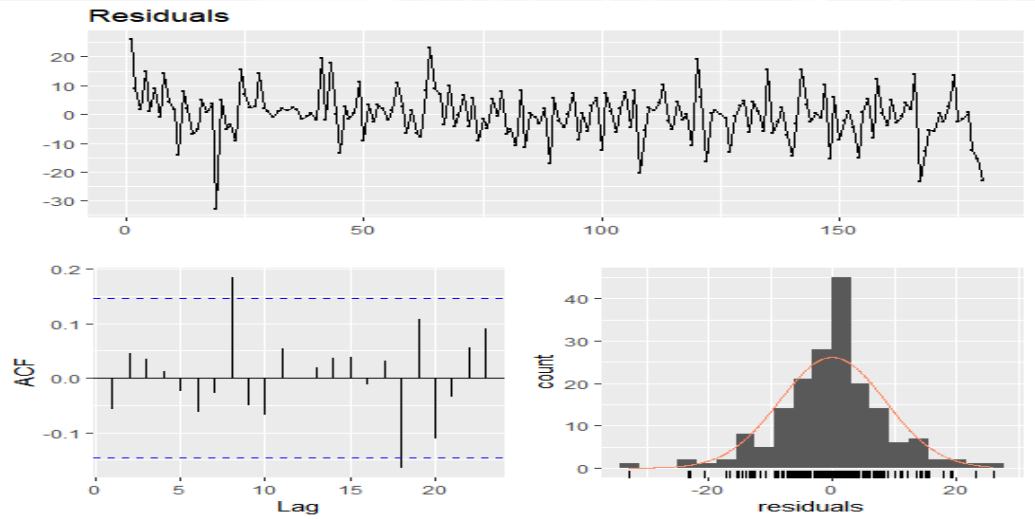
▶ 그래프로 볼 때, 정규성은 만족한 것으로 보인다.

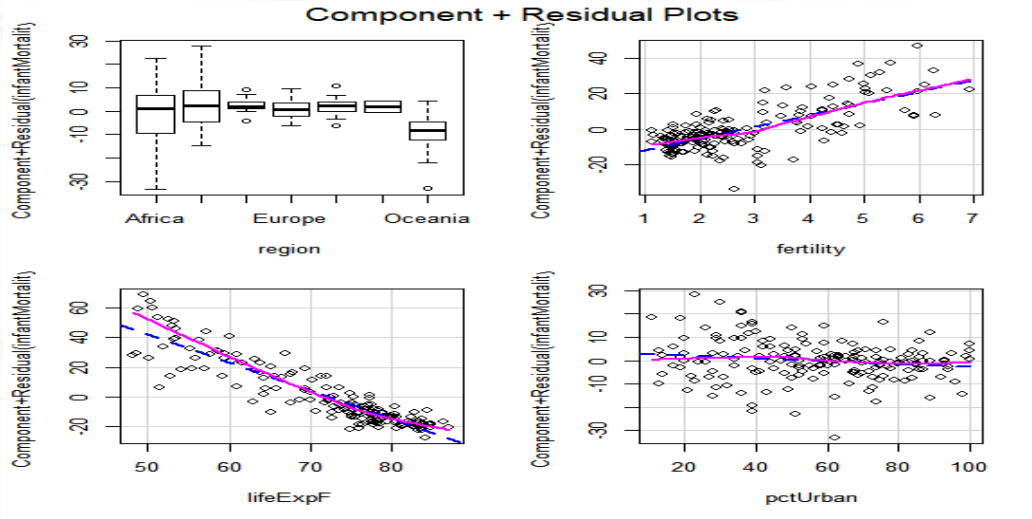
04 잠정 모형 확인

세 번째 잠정 모형

- $Y_{infantMortality} = 152.07 + 2.40\, D_{regionAsia} + \cdots - 9.16\, D_{regionOceania} + 6.56\, X_{fertility} - 1.89\, X_{lifeExpF} - 0.06\, X_{pctUrban}$

- 유의수준 10% 하에서, Intercept, region, fertility, lifeExpF, pctUrban 모든 변수가 유의적





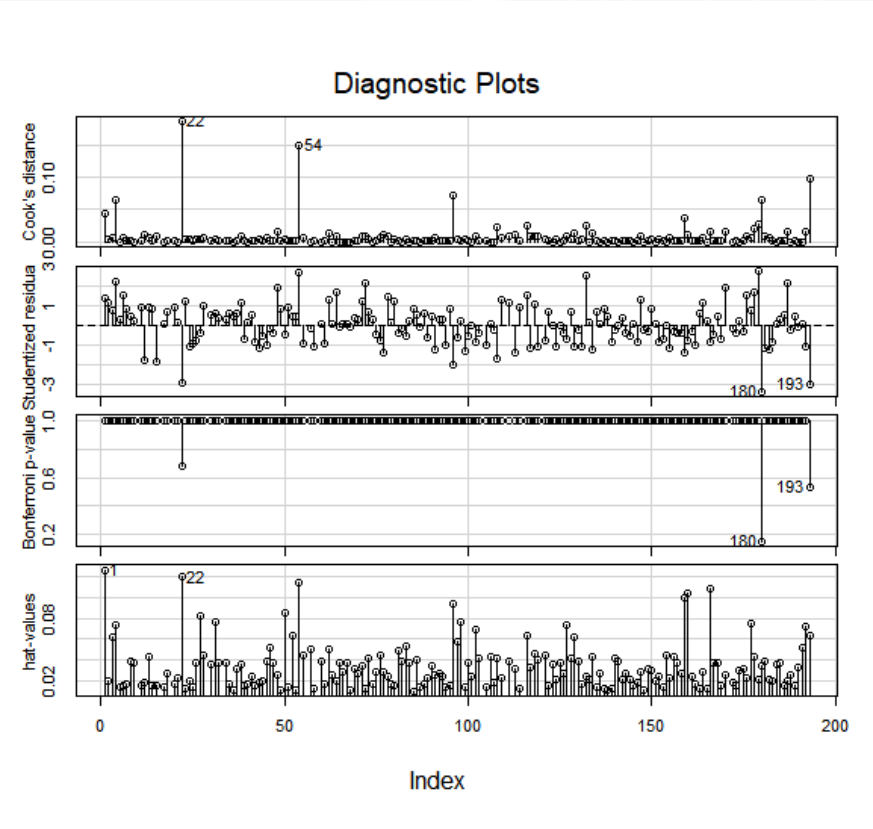
▶ Breusch–Godfrey test 결과 p–value가 0.5031 로 독립성은 만족하는 것으로 보인다.

▶ 반응 변수와 설명 변수간의 선형 관계 역시 만족하지만, pctUrban은 큰 영향력이 없어 보인다.

04 잠정 모형 확인

세 번째 잠정 모형의 특이한 관찰값 탐지

- $Y_{infantMortality} = 152.07 + 2.40\, D_{regionAsia} + \cdots - 9.16\, D_{regionOceania} + 6.56\, X_{fertility} - 1.89\, X_{lifeExpF} - 0.06\, X_{pctUrban}$



- 특이한 관찰값 후보 : Afghanistan(1), Botswana(22), Tuvalu(180)

- 특이한 관찰값 제거 후 회귀 모형 적합

▶ Intercept, region, fertility, lifeExpF, pctUrban 모든 변수가 유의적

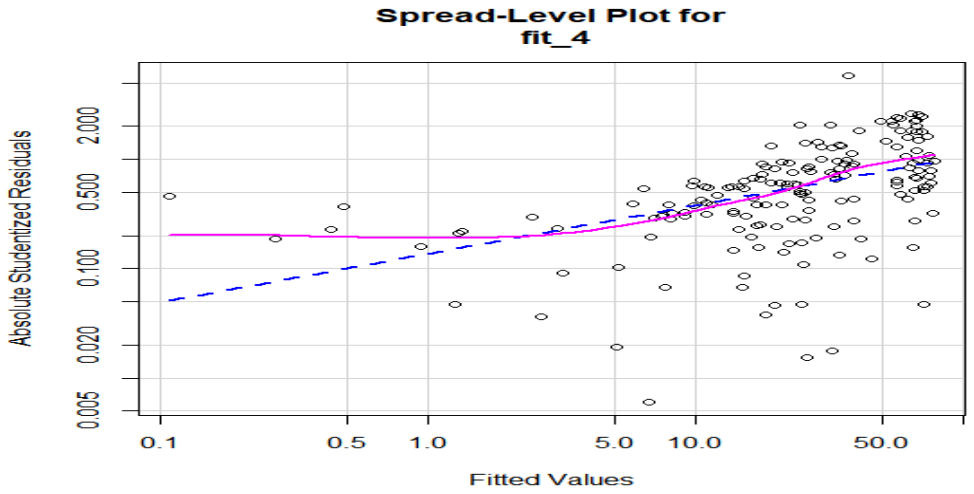
▶ 등분산, 정규성, 독립성, 선형 관계 모두 만족

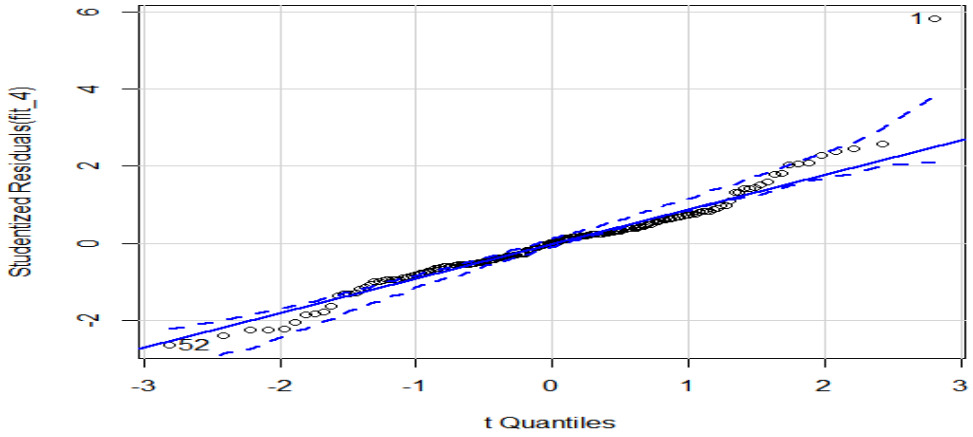
04 잠정 모형 확인

네 번째 잠정 모형

- $Y_{infantMortality} = 41.36 + 5.62\, D_{groupother} + 41.55\, D_{groupafrica} - 0.0001\, X_{ppgdp} - 0.41\, X_{pctUrban}$

- 유의수준 10% 하에서, Intercept, group, pctUrban 변수가 유의적. ppgdp는 비유의적





▶ 분산이 점점 증가는 것으로 보인다. → 반응 변수의 변환

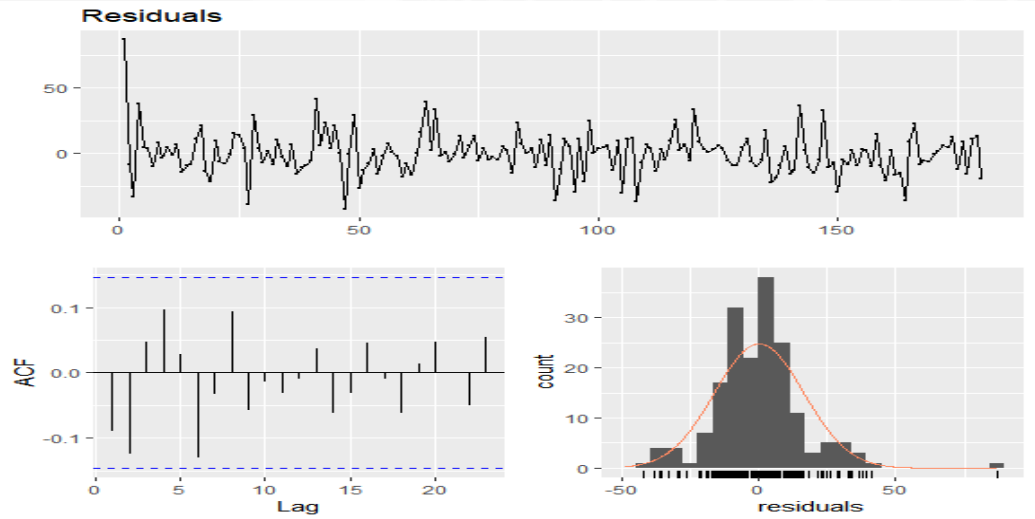
▶ 그래프로 볼 때, 정규성은 만족한 것으로 보인다.

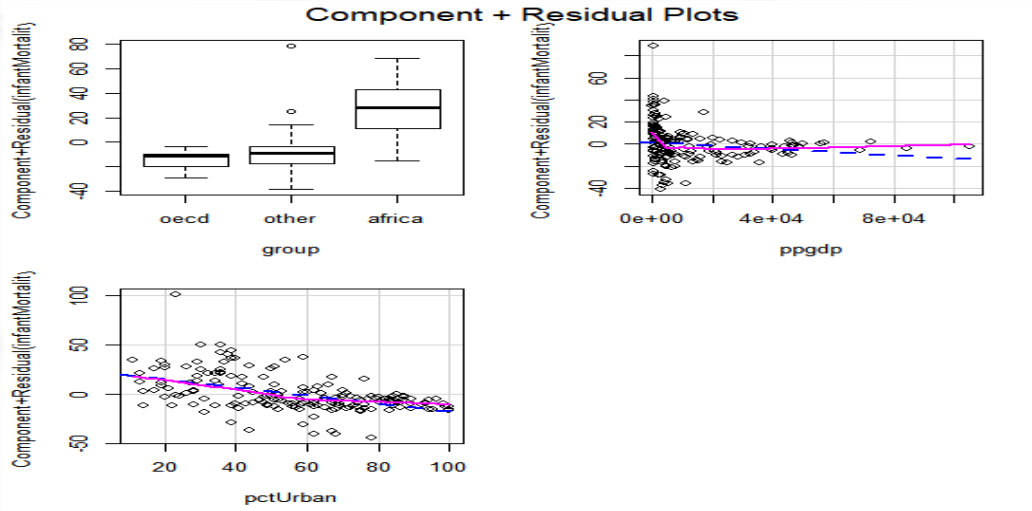
04 잠정 모형 확인

네 번째 잠정 모형

- $Y_{infantMortality} = 41.36 + 5.62\, D_{groupother} + 41.55\, D_{groupafrica} - 0.0001\, X_{ppgdp} - 0.41\, X_{pctUrban}$

- 유의수준 10% 하에서, Intercept, group, pctUrban 변수가 유의적. ppgdp는 비유의적





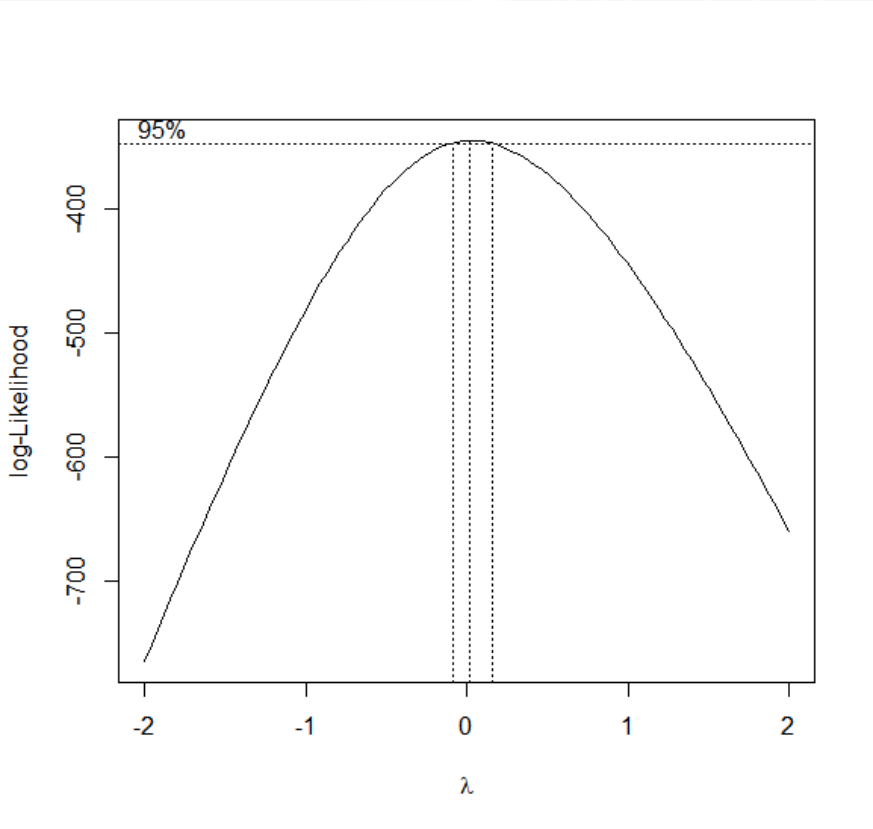
▶ Breusch–Godfrey test 결과 p–value가 0.4234 로 독립성은 만족하는 것으로 보인다.

▶ 반응 변수와 설명 변수간의 선형 관계 역시 만족하지만, ppgdp, pctUrban은 큰 영향력이 없어 보인다.

04 잠정 모형 확인

네 번째 잠정 모형의 반응 변수 변환

• 동일 분산 가정을 만족시키기 위해 반응 변수의 변환이 필요하다.



• 추정된 $\hat{\lambda} = 0.02$

• $\hat{\lambda} = 1$ 에 대한 가설 검정 : 귀무가설 기각. 변수 변환 필요

• $\hat{\lambda} = 0$ 에 대한 가설 검정 : 귀무가설 채택. 로그 변환 실시

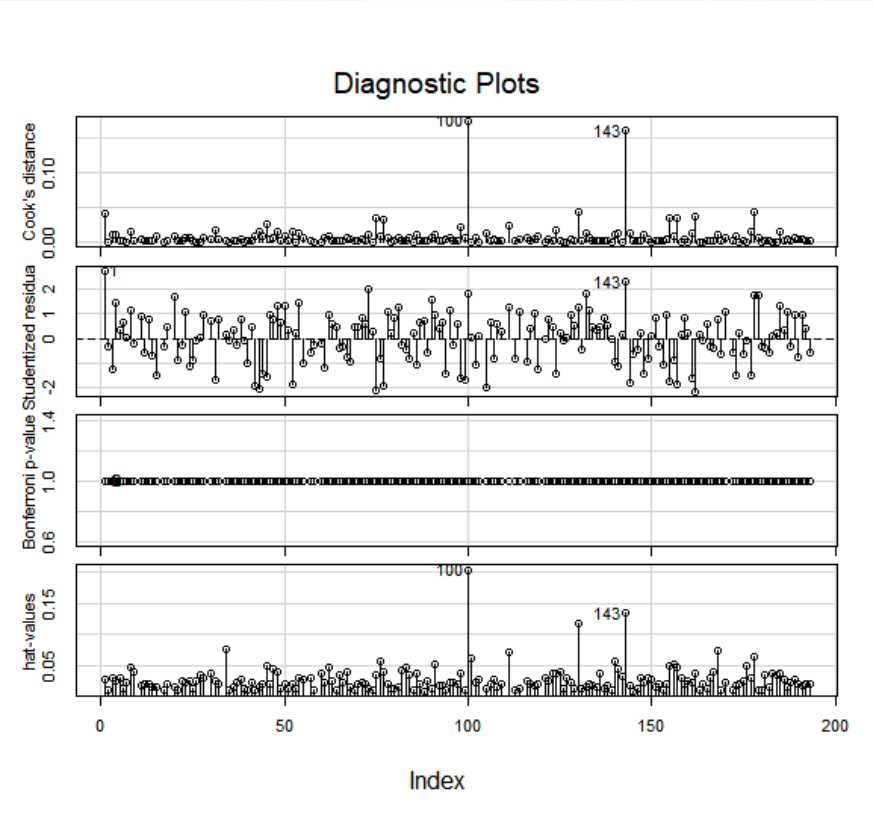
▶ $\hat{\lambda} = 0$ 즉, 로그(log) 변환 실시.

▶ $y \rightarrow \log(y)$

04 잠정 모형 확인

네 번째 잠정 모형의 특이한 관찰값 탐지

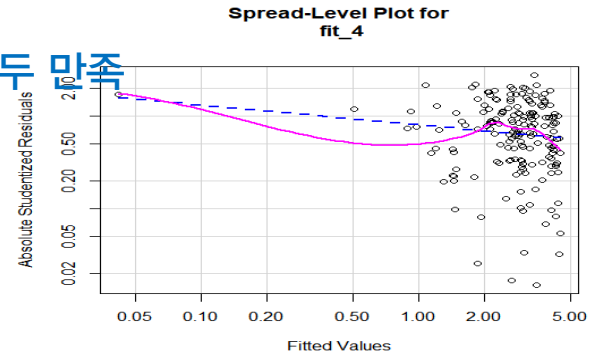
- $\log(Y_{infantMortality}) = 41.36 + 5.62\, D_{groupother} + 41.55\, D_{groupafrica} - 0.0001\, X_{ppgdp} - 0.41\, X_{pctUrban}$



- 특이한 관찰값 후보 : Luxembourg (100), Qatar(143)
- 특이한 관찰값 제거 후 회귀 모형 적합

▶ 절편, OECD, GDP, 도시 비율 모든 변수가 유의적

▶ 등분산, 정규성, 독립성, 선형 관계 모두 만족



04 잠정 모형 확인

잠정 모형

- 반응 변수 변환과 특이한 관찰값 제거로 모든 잠정 모형이 가정 만족.

잠정 모형 1

- $\sqrt{Y_{infantMortality}} = 133.13 + 4.68\,X_{fertility} - 1.75\,X_{lifeExpF} + 120.11\,X_{ppgdp\text{^}-0.3}$

잠정 모형 2

- $\sqrt{Y_{infantMortality}} = 140.08 - 4.02\,D_{groupother} - 4.10\,X_{groupafrica} + 4.44\,X_{fertility} - 1.80\,X_{lifeExpF} + 131.65\,X_{ppgdp^{-0.3}}$

잠정 모형 3

- $Y_{infantMortality} = 152.07 + 2.40\,D_{regionAsia} + \cdots - 9.16\,D_{regionOceania} + 6.56\,X_{fertility} - 1.89\,X_{lifeExpF} - 0.06\,X_{pctUrban}$

잠정 모형 4

- $\log(Y_{infantMortality}) = 41.36 + 5.62\,D_{gorupother} + 41.55\,D_{gorupafrica} - 0.0001\,X_{ppgdp} - 0.41\,X_{pctUrban}$

05.

최종 모델 선택

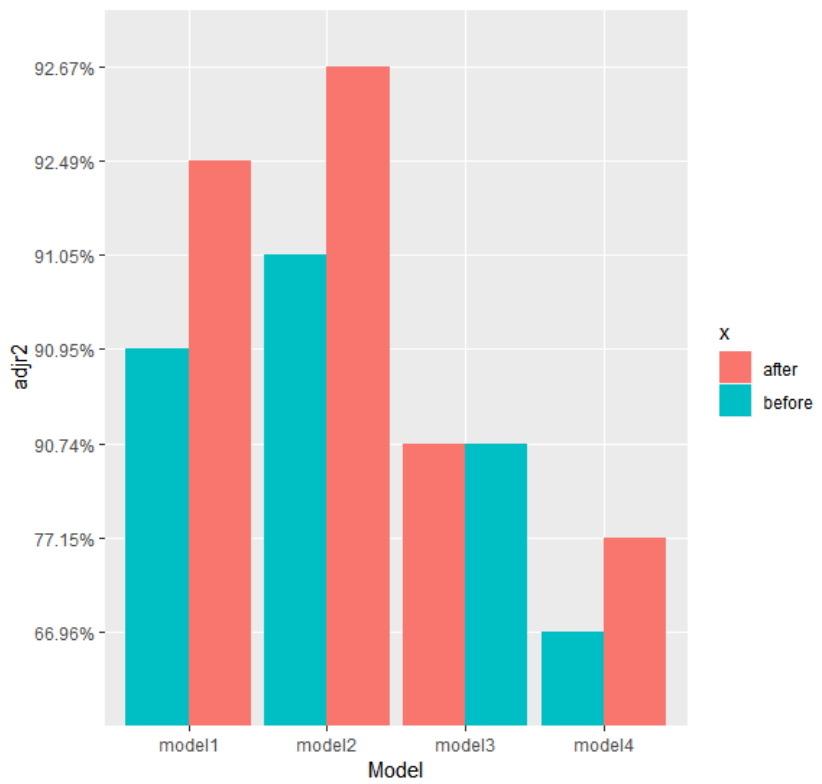
Final Model

05 최종 모델 선택 및 예측

모형 선택 기준

• 수정된 결정 계수 : adj R^2, 모형에 의해 설명되는 반응 내 변동의 백분위

반응변수 변환과 이상값 제거 전후의 수정된 결정 계수



- model1 : 90.95% → 92.49% (1.54% 상승)
- model2 : 91.05% → 92.67% (1.62% 상승)
- model3 : 90.74% → 90.74% (0% 상승)
- model4 : 66.96% → 77.15% (10.19% 상승)

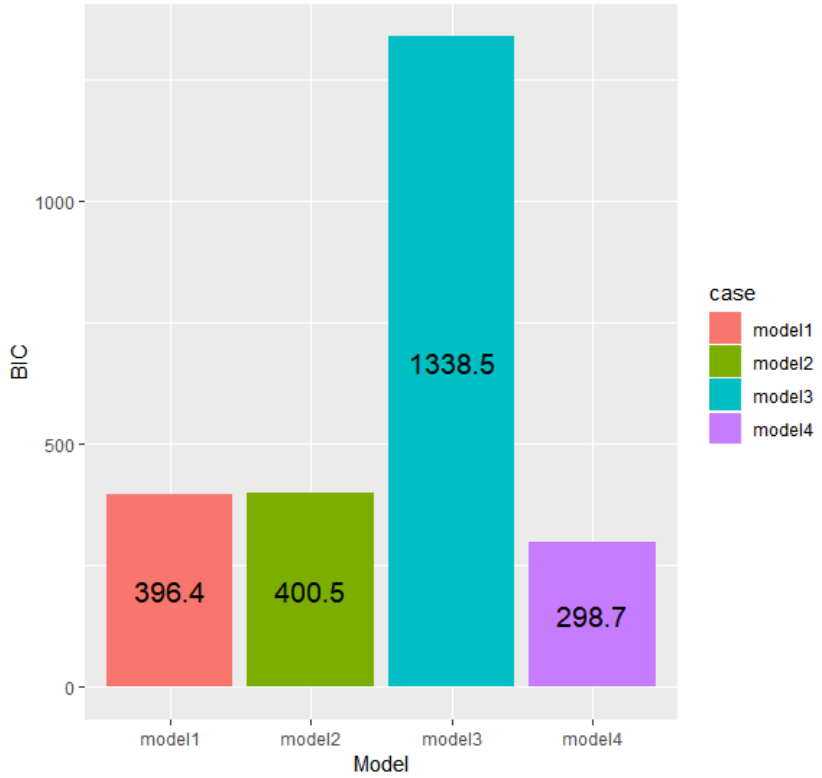
▶ 수정된 결정 계수가 가장 높은 Model 2 을 선택

05 최종 모델 선택 및 예측

모형 선택 기준

- BIC : $-2\log L + K\log n$, 값이 작을수록 올바른 모형에 가깝다.

BIC



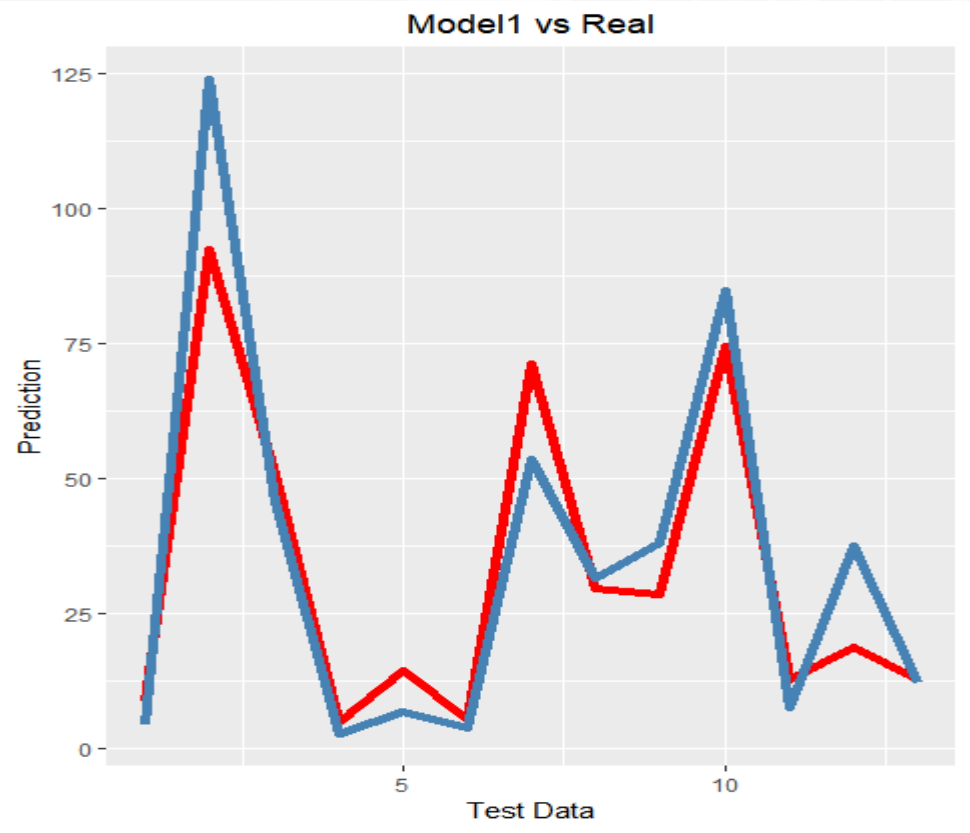
- model1 : 396.4
- model2 : 400.5
- model3 : 1338.5
- model4 : 298.7

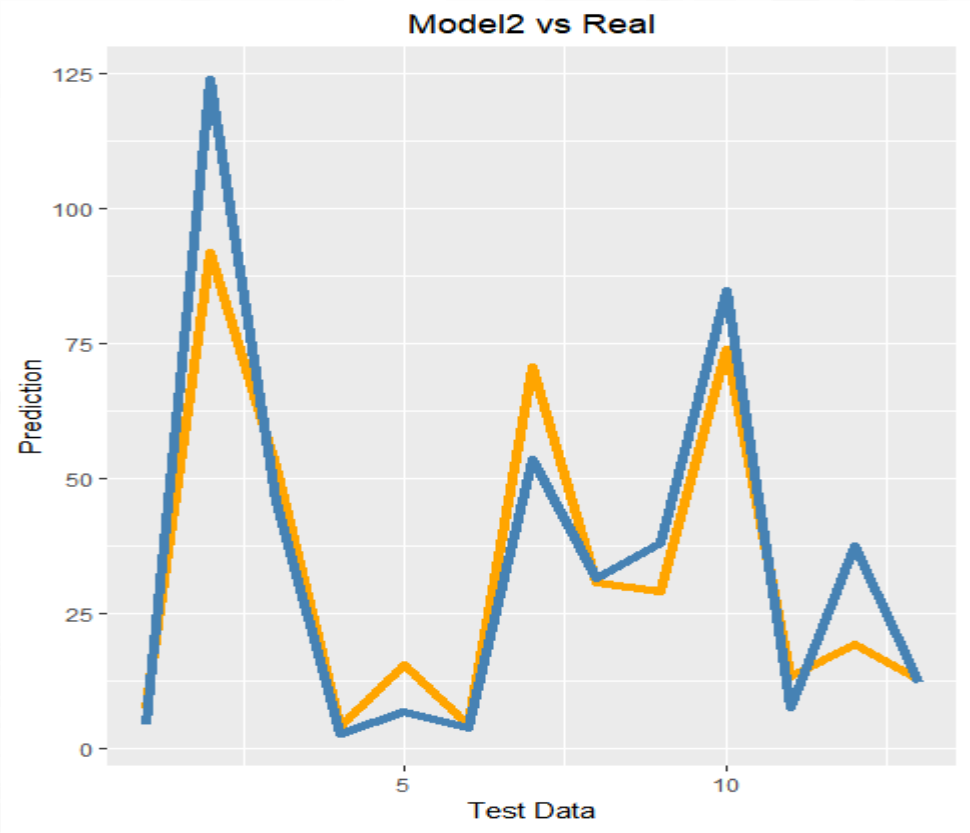
▶ BIC 값이 가장 낮은 Model 4 을 선택

05 최종 모델 선택 및 예측

Test data 예측

- 13개의 Test data에 대한 각 Model 들의 예측 및 Accuracy 확인

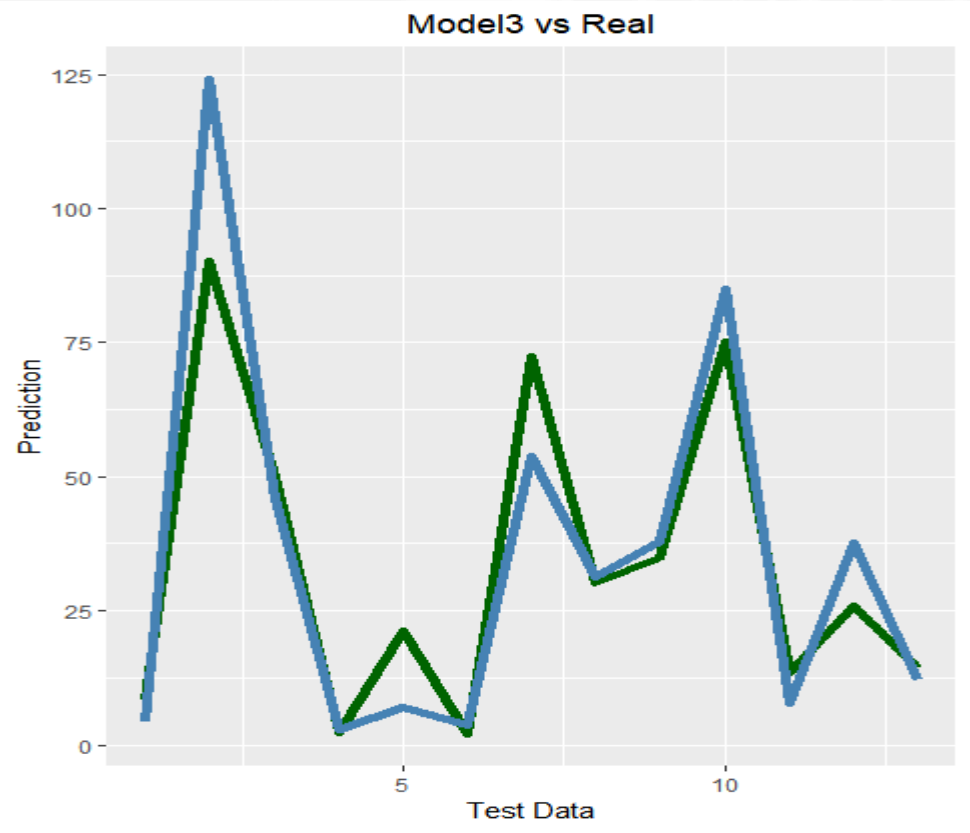




05 최종 모델 선택 및 예측

Test data 예측

- 13개의 Test data에 대한 각 Model 들의 예측 및 Accuracy 확인



05 최종 모델 선택 및 예측

Test data 예측

- 13개의 Test data에 대한 각 Model 들의 예측 및 Accuracy 확인

- 빨간선 : Model 1
- 주황선 : Model 2
- 노랑선 : Model 3
- 초록선 : Model 4
- 파랑선 : 실제값

05 최종 모델 선택 및 예측

Test data 예측

- 13개의 Test data에 대한 각 Model 들의 예측 및 Accuracy 확인

Accuracy 비교

	ME	RMSE	MAE	MPE	MAPE
Model 1	−2.16	12.40	8.94	5.06	34.18
Model 2	−2.18	12.40	8.73	3.26	30.44
Model 3	−0.98	12.43	8.52	0.43	31.50
Model 4	−5.81	24.72	18.30	−31.57	72.85

▶ Scale에 영향을 받는 지표 : MAE, RMSE

▶ Scale에 영향을 받지 않는 지표 : MAPE

▶ RMSE, MAPE 가 낮은 Model 2 가 가장 좋은 모델

05 최종 모델 선택 및 예측

최종 모형

• Model 2 : $\sqrt{Y_{infantMortality}} = 140.08 - 4.02\, D_{groupother} - 4.10\, D_{groupafrica} + 4.44\, X_{fertility} - 1.80\, X_{lifeExpF} + 131.65\, X_{ppgdp^{-0.3}}$

• RMSE : 12.40 • MAPE : 30.44

• adj R^2 : 92.67

• BIC : 400.5

▶ 수정된 결정 계수 값이 92.67 로 가장 높음.

▶ RMSE, MAPE 값이 가장 낮음.

▶ 그래프 역시 Test 데이터를 다른 모델에 비해 잘 예측함.

06.

분석 결론

Result of Analysis

06 분석 결론

UN 데이터란?

2009년 ~ 2011년 사이의 UN에 가입되어있는 국가들의 통계자료를 기록한 데이터이다.

• 각 나라별 빈부격차의 심화

- group(OECD) 범주에 따라 fertility, lifeExpF, ppgdp, pctUrban, infantMortality의 값의 차이가 심했다.
- OECD 가입국과 Africa 국가들의 영유아 사망률의 차이가 심했다.

• 영유아 사망률에 가장 큰 영향을 미친 변수는 무엇인가?

- 최종 모델(Model 2) 의 경우, OECD가입 여부, 출산율, 여성 수명, GDP 가 포함된 모델이다.
- ANOVA table 에서 Sum of Sq 가 가장 큰 값은 group이고, fertility, lifeExpF, $ppgdp^{-0.3}$ 순이다.

06 분석 결론

UN 데이터란?

2009년 ~ 2011년 사이의 UN에 가입되어있는 국가들의 통계자료를 기록한 데이터이다.

• 분석 보완점

- 각 group 별로 회귀 모델을 세우는 것이 더 정확한 예측을 할 수 있을 것 같다.
- 각 변수 마다 Scale 의 차이가 심하기 때문에 표준화 등의 데이터 전처리 과정이 필요해 보인다.
- 회귀 모델의 일반화를 위해 교차 검증(Cross Validation)을 통한 모델 평가가 필요해 보인다.

감사합니다.

2조 박상희 | 이인풍 | 김문경 | 정중한